SHRINKAGE ESTIMATION FOR PENALISED REGRESSION, LOSS ESTIMATION AND TOPICS ON LARGEST EIGENVALUE DISTRIBUTIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Rajendran Narayanan

August 2012

SHRINKAGE ESTIMATION FOR PENALISED REGRESSION, LOSS ESTIMATION AND TOPICS ON LARGEST EIGENVALUE DISTRIBUTIONS

Rajendran Narayanan, Ph.D.

Cornell University 2012

The dissertation can be broadly classified into four projects. They are presented in four different chapters as (a) Stein estimation for l_1 penalised regression and model selection, (b) Loss estimation for model selection, (c) Largest eigenvalue distributions of random matrices, and (d) Maximum domain of attraction of Tracy-Widom Distribution.

In the first project, we construct Stein-type shrinkage estimators for the coefficients of a linear model, based on a convex combination of the Lasso and the least squares estimator. Since the Lasso constraint set is a closed and bounded polyhedron (a crosspolytope), we observe that under a general quadratic loss function, we can treat the Lasso solution as a metric projection of the least squares estimator onto the constraint set. We derive analytical expressions for the decision theoretic risk difference of the proposed Stein-type estimators and Lasso and establish data-based verifiable conditions for risk gains of the proposed estimator over Lasso. Following the Stein's Unbiased Risk Estimation (SURE) framework, we further derive expressions for unbiased esimates of prediction error for selecting the optimal tuning parameter.

In the second project, we consider the following problem. For a random vector X, estimation

of the unknown location parameter θ using an estimator $d(X)$ is often accompanied by a loss function $L(d(X), \theta)$. Performance of such an estimator is usually evaluated using the risk of $d(X)$. We consider estimating the loss function using an estimator $\lambda(X)$ which is conditional on the actual observations as opposed to an average over the sampling distribution of $d(X)$. In this context, we consider estimating the loss function when the unknown mean vector θ of a multivariate normal distribution with an arbitrary covariance matrix is estimated using both the MLE and a shrinkage estimator. We derive sufficient conditions for inadmissibility of the unbiased estimators of loss for such a random vector. We further establish conditions for improved estimators of the loss function for a linear model when the Lasso is used as a model selection tool and exhibit such an improved estimator.

The largest eigenvalue of the Gaussian and Jacobi ensembles plays an important role in classical multivariate analysis and random matrix theory. Historically, the exact distribution for the largest eigenvalue has required extensive tables or use of specialised software. More recently, asymptotic approximations for the cumulative distribution function of the largest eigenvalue in both settings have been shown to have the Tracy-Widom limit. Our main results concern using a unified approach to derive the exact cumulative distribution function of the largest eigenvalue in both settings in terms of elements of a matrix that have explicit scalar analytical forms.

In the fourth chapter, the maximum of i.i.d. Tracy-Widom distributed random variables arising from the Gaussian unitary ensemble is shown to belong to the Gumbel domain of attraction. This theoretical result has potential applications in any situation where a multiple comparisons is needed using the greatest root statistic.

BIOGRAPHICAL SKETCH

Rajendran Narayanan spent his formative years in the city of Kolkata, India and finished his higher secondary schooling from St. Lawrence High School, Kolkata. Upon completion, he obtained his Bachelors degree in Statistics (Honours), Mathematics and Economics from St. Xavier's College, Kolkata in 1999. He then completed his Masters degree in Applied Statistics and Informatics from the Department of Mathematics at the Indian Institute of Technology, Mumbai, India in 2002.

After two years of working in the industry as a Business Analyst, he swtiched to applied industrial research working as a Scientist with the GE Global Research Centre, Bangalore, India. Following which, Rajendran began his graduate studies in Statistics at the Department of Statistical Science at Cornell University, Ithaca, U.S.A. in 2006. Upon completion of his PhD in August 2012, he would be joining the Indian Statistical Institute in Kolkata as a Visiting Scientist.

Dedicated to my mother and to the memory of my high school teacher, Shri Lal Bahadur Singh.

ACKNOWLEDGEMENTS

There have been numerous people in several capacities who have been very instrumental during my PhD life. I have been very fortunate to have Prof. Martin Wells as my thesis supervisor. He has been great at articulating my intuition into a rigorous framework and helping me concretise the right questions to ask. As a research mentor, I particularly thank him for the ample time and free rein he has given me to explore various areas of mathematical statistics in addition to retaining humour and exercising patience during my ebb times in research. With his unbridled optimism and infectious enthusiasm, he has nurtured and actively supported my quest for a holistic development by indulging me in several of my non academic pursuits. Suffice it to say that I have learned a lot about life well beyond the domain of research through my discussions with him on a myriad of subjects.

I wish to express my sincere thanks to my committee members, Prof. Rob Strawderman and Prof. Michael Nussbaum. Rob's keen insight and great attention to detail has played a crucial role in this thesis. In particular, he pointed a flaw in my proof on the domain of attraction of the Tracy-Widom distribution and later helped me tide over a key step in the new proof of the theorem. I sincerely thank him for exposing me to inference of stochastic processes and the dedication he has displayed towards my learning at graduate school. My special thanks to Michael for directing me to the literature on empirical risk minimisation and helping me formalise my geometric ideas of embedding lasso in a cone.

I have been lucky to have had the opportunity to present and discuss my thoughts on the first chapter of the thesis with Prof. Dominique Fourdrinier who had valuabale comments on how to cogently structure the material in the chapter.

Special thanks are in order for Diana Drake. She has been an indispensable source of support; from rescuing me from all the problems I usually landed up in school (owing to my tomfoolery and innate procrastinating self) to bringing back to life my quasi dead car two days before Christmas. None of this work would have been possible without the unseen support staff of Malott Hall who provided a clean and aesthetic environment to work.

Surviving the course work phase wouldn't have been possible without Caitlin and Kirsten. Dave was particularly helpful in dusting off my rusted concepts of real analysis and enhancing my mathematical experience in general. Thanks to Michael Grabchak for being the sceptical inquirer of my half-baked proofs in addition to presenting the Marxist (of the Groucho type) perspective during our endless debates about politics, society and cinema over innumerable pints. Several hours over coffee and beer were spent on philosophical excursions, self reflection and jollity with Matti. Apart from being the sounding board on some eigenvalue proofs, the monk-like Inder has been a close comrade discussing with me, everything from the sublime to the ridiculous. Anand, Suresh and Vivek were phenomenally easy-to-live-with house-mates who not only bore the brunt of my "Cries and Whispers" but also provided great fodder for scientific thinking. The unyielding courage and passion for the quest for "truth" displayed by Vatsan and Saikat will remain a learning experience for life. Leifur, Darcy, Mekala and Ryan have been fantastic friends who've lent patient ears to many of my rants and ramblings. Thanks to Jon for being ever ready for badminton and beer. The positive spirit of Matt, Krishna's benevolent attitude and Joyjit's encouraging words have also played a key role during this phase. The last lap of this journey has been extremely easy due to the witty, calm and tolerant Kalyani whose ability to handle my neurotic self

borders on quixotic lines. My heartfelt acknowledgements to Chapter House for providing the platform for my interactions with most of the people mentioned in this paragraph.

I am very grateful to Venu and my sister Shobana to have given me a lot of support and having extended unparalleled warmth during my breaks to their house. Finally, I would like to thank my mother who is not only a genuinely liberal friend but also a massive source of inspiration who has taught me to think independently.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Shrinkage Estimation for l_1 Penalised Regression

1.1 Introduction

There has recently been considerable attention on applying shrinkage ideas to model selection and parameter estimation in regression models. Compared with ordinary least squares (OLS), shrinkage methods often improve the prediction accuracy and, if a defined constraint set has edges or corners, some parameter estimates can be shrunk exactly to zero. A very popular method in this regard is the Lasso, proposed by Tibshirani [45]. Its popularity stems from the fact that it serves the dual role of model selection and estimation. In particular, when the number of variables is large, a common situation in genomics, astronomy and meteorological data among other fields, the problem of variable selection is combinatorially hard. The Lasso is a very effective method in such situations. It is obtained by minimising the error sum of squares subject to an l_1 norm constraint on the regression coefficients. In other words, the Lasso estimate, $\hat{\beta}_t$, is obtained as

$$\hat{\beta}_t = \underset{\beta \in L_t}{\operatorname{argmin}} \|y - X\beta\|_2^2 \tag{1.1}$$

where $L_t = \{\beta \in \mathbb{R}^p : \sum_{j=1}^p |\beta_j| \leq t\}$. The equivalent Lagrange multiplier formulation is given by

$$\hat{\beta}_t = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta_j| \tag{1.2}$$

where t and λ are nonnegative tuning parameters.

The solutions to (1.1) and (1.2) are sparse, thus making the Lasso popular among practitioners as a model selection procedure. There has been increasing interest in the sparse solution's theoretical properties. Knight and Fu [28] provide consistency results for Lasso type problems under some regularity conditions on the design matrix. They further provide the limiting distribution of the Lasso solution. Zhao and Yu [57] provide a condition which serves as a necessary and sufficient condition for the Lasso to select the true model. More recently, Chatterjee and Lahiri [8] propose a modified bootstrap method to provide approximations of the distribution of the Lasso estimator. They also prove consistency in estimating the asymptotic bias and variance. There are numerous papers establishing various other theoretical properties and extensions in various other settings. Candes and Tao [7] examine the restricted isometry property that serves as a condition for the coding matrix to satisfy to recover a sparse and corrupted signal. Their result focuses on the uniform Gaussian ensemble in the noiseless setting and an under-determined system ($n = \gamma p$ for some $\gamma \in (0, 1)$). They prove that there exists some $\alpha > 0$ such that all sparsity patterns with $s \leq \alpha p$ can be recovered with high probability in the asymptotic setting. Wainwright [51] gives a necessary and sufficient condition that the sample size n, the problem dimension p, and the sparsity s have to satisfy for recovery of the true sparsity with high probability. In addition, he also provides the thresholds based on the eigenvalues of a random Gaussian design matrix to recover the true sparsity with a high probability.

Shrinkage estimation has, arguably, been one of the most active research areas in statistics over the last several decades. For a random vector X having a multivariate normal distribution, $N_p(\theta, I_p)$, the usual estimator X is the uniform minimum variance unbiased estimator (UMVUE), best equivariant location estimator as well as the maximum likelihood estimator

(MLE) of the mean vector. It has a constant risk and is also a minimax estimator. However, when the dimension of the parameter space exceeds two, the MLE becomes inadmissible. James and Stein [22] proved that the, now classical, estimator $(1 - a\|X\|^{-2})X$ dominates the MLE for $0 < a < 2(p-2)$ and that $a = p - 2$ gives the uniformly best estimator in this class.

One of the many possible extensions and generalisations of the canonical setting of Stein estimation is when the parameter space is restricted. An early work in this area was done by Bock [5] who proposed Stein-like estimators for the multivariate normal distribution with identity covariance matrix when the mean vector is constrained to lie in a closed and convex polyhedron. A remarkable fact proved there is that the domination of the Stein-type estimator over the MLE was dependent on the *codimension* of the face of projection of the MLE onto the constraint set and not just on the ambient dimension. Sengupta and Sen [40] considered the case when the parameter space is a positively homogeneous cone. Since Stein type estimators dominate the MLE, they construct Stein-type estimators that dominate the restricted MLE under various forms of conical restrictions. A further extension of this idea, proposed by Kuriki and Takemura [30] developed shrinkage estimators when the parameter space was constrained to be a closed and convex set with a smooth or piecewise smooth boundary. Their estimators are derived by considering a sequence of polytopes converging to the constraint set and looking at the limit of the corresponding shrinkage estimators to the polytopes. Bock's estimator turns out to be a special case of this result when the constraint set is polyhedral.

From an asymptotic view there is a body of literature devoted to identifying adaptive min-

imax estimators for various estimation problems, which are simultaneously asymptotically minimax over a range of restricted parameter spaces [e.g. q-balls] as in Donoho [13], and Donono and Jonstone [14]. The present work has close connections to the normal means problem and minimax estimation over the 1-ball. A key observation in our exercise has been the role of sparsity in risk domination. A basic result in this direction is that the James-Stein estimator for the normal means problem is adaptive asymptotic minimax over 2-balls, see for example [14], but not asymptotically minimax over q-balls for $q < 2$ [13]. Minimax estimators for q-balls, with $q < 2$, generally perform well when the parameter of interest is sparse and, for the normal means problem, thresholding estimators were introduced to achieve asymptotic minimaxity in this setting. We observe both in simulation and real data analysis that the extent of sparsity in the data has a direct bearing in the amount of risk dominance noticed.

In the estimator we consider, the shrinkage factor lies between 0 and 1. Our proposed estimator becomes the Lasso solution when the shrinkage factor equals one. Lasso thus acts as a boundary case for the proposed shrinkage estimator. This is also corroborated in simulations when we observe that when the true number of nonzero coefficients proportional to the total number of variables in a model is large, the shrinkage factor is closer to 0 and we note large risk gains of the shrinkage estimator over the Lasso. However, when the true number of nonzero coefficients as a proportion of the total number of variables in the model is few, the adaptive shrinkage factor is closer to one and we will see that the decision theoretic risks of the shrinkage model and the Lasso are comparable.

On the computational side, Osborne et al. [35] presented an efficient algorithm to compute

the Lasso solutions treating the problem as a convex programming problem. The seminal paper by Efron et al. [17] on least angle regression (LARS) provides a robust computational algorithm to compute the entire Lasso solution path with the running time of a single least squares fit. The latter algorithm also provides C_p as an output of the fitting process as a tool for selecting the optimal tuning parameter for model selection.

We treat the Lasso as a projection of the least squares estimator onto the constraint set L_t; geometrically a p-crosspolytope, and propose shrinkage estimators of the regression coefficients in the linear model setting. In this setting, we can also view the Lasso as the restricted MLE on L_t. We then establish closed form analytical formula of the risk difference between the proposed shrinkage estimator and Lasso and give conditions for dominance of the proposed estimator over the Lasso. In addition, we also give data based closed-form expressions for estimates of prediction risk. Opting for a more conservative model selection approach, we use the smallest prediction risk as the criterion for choosing the optimal shrinkage estimator.

In Section 1.2 we give the results pertaining to estimation of risk of the regression parameters and establish conditions of dominance over Lasso and the results are proved in Section 1.6. In Section 1.4 we derive explicit analytical expressions for unbiased estimates of the prediction risk which are important for model choice tuning parameter selection. In Section 1.5 we present the simulation methodology followed by a presentation of an idea on ensuring sparsity using our approach in Section 1.3.1. Next we present some simulation results and some results of analysis performed on real data, we use the diabetes data for this purpose and present a comparison with the Lasso. Tables and Figures are presented in Section 1.8.

1.2 Shrinkage Estimators

Throughout the chapter, we denote $\hat{\beta}_0$ as the ordinary least squares (OLS) estimator and $\hat{\beta}_t$ as the Lasso estimator for any arbitrary tuning parameter. Under a general quadratic norm, the Lasso is a metric projection of the least squares estimator on the compact and convex constraint set. As such, we also refer to the Lasso as the projection estimator. Using the least squares estimator and the projection estimator, we can construct an improved (a Stein-type) estimator for the coefficients of the linear model as

$$\hat{\beta}_\phi = \hat{\beta}_t + (1 - \phi(\cdot))(\hat{\beta}_0 - \hat{\beta}_t), \tag{1.3}$$

where $\phi(\cdot)$ is a real-valued function called the shrinkage factor, such that $0 < \phi(\cdot) < 1$. The natural question is why should the above equation be called the improved estimator. We have imposed the Lasso constraint on the regression coefficients, i.e. , $\beta \in L_t$. The first term in the right hand side of (1.3) refers to the projection of $\hat{\beta}_0$ onto L_t. This term captures the maximum possible information about the unknown coefficients given the OLS estimates. In other words, because $\hat{\beta}_t \in L_t$, it corresponds to the component of the observation that conforms to the prior belief indicating L_t as the set in which a reasonable estimate should lie. On the other hand, $(\hat{\beta}_0 - \hat{\beta}_t)$ corresponds to the deviation of the observation from the prior belief. As such, we would like to shrink this component to get improved estimators. We get the following as boundary cases: $\phi(\cdot) = 0$ gives $\hat{\beta}_0$, the least squares estimator indicating no shrinkage and $\phi(\cdot) = 1$ gives $\hat{\beta}_t$, the projection estimator or Lasso indicating complete shrinkage.

To choose the shrinkage factor, $\phi(\cdot)$, we use the results of Bock [5], and Kuriki and Takemura [30], and set the shrinkage factor to be $\phi(l) = (m-2)/l^2$ where m is the codimension of the

face of projection of the least squares solution on the polytope and l^2 denotes the squared length of projection between the least squares solution and the Lasso. Observe that, the results in [5] and [30] treat the constraint set as fixed, in the sense that they are not indexed by a tuning parameter. However, in the case we consider, the constraint set L_t is indexed by a tuning parameter and owing to the nature of the algorithms performing Lasso, the codimension of the face of projection, m and the squared length of projection, l^2 changes for each value of t. Consequently, what we obtain is a sequence of shrinkage estimators indexed by t as opposed to a fixed shrinkage estimator proposed in [5] and [30]. Hence the choice of t for optimal estimator selection is important, addressed in Section 1.4.

Define the difference in loss functions due to Lasso and OLS as

$$\Delta L_{10} = \|\hat{\beta}_t - \beta\|_{\Sigma}^2 - \|\hat{\beta}_0 - \beta\|_{\Sigma}^2.$$

Define the difference in loss functions due to $\hat{\beta}_\phi$ and OLS as

$$\Delta L_{\phi 0} = \|\hat{\beta}_\phi - \beta\|_{\Sigma}^2 - \|\hat{\beta}_0 - \beta\|_{\Sigma}^2.$$

$\Delta L = \Delta L_{\phi 0} - \Delta L_{10}$ gives the difference in the loss functions of $\hat{\beta}_\phi$ and the Lasso. Let ΔR denote the difference in risk between $\hat{\beta}_\phi$ and the Lasso.

$$\begin{aligned}
\Delta R &= E_\beta[\Delta L] \\
&= E_\beta[\Delta L_{\phi 0}] - E_\beta[\Delta L_{10}] \\
&= \Delta R_{\phi 0} - \Delta R_{10}.
\end{aligned}$$

where $\Delta R_{\phi 0}$ denotes the risk difference between $\hat{\beta}_\phi$ and the least squares estimator and ΔR_{10} denotes the risk difference between the Lasso and least squares estimator.

1.2.1 Main Results

We state the main results of this chapter below and give their proofs in Section 1.6. Theorem 1.1 provides an unbiased estimator of the risk difference between the Lasso solution and the least squares solution, i.e., it gives an expression for $\widehat{\Delta R_{10}}$ such that $E_\beta[\widehat{\Delta R_{10}}] = \Delta R_{10}$. In what follows, we consider the standard linear model $Y = X\beta + \epsilon$ where X is a design matrix having n observations on p variables and $\epsilon \sim N(0, \sigma^2)$, independent of X. It is well known that if σ^2 is known, then $\hat{\beta}_0 \sim N(\beta, \sigma^2(X^TX)^{-1})$. We denote $\Sigma^{-1} = \sigma^2(X^TX)^{-1}$.

Theorem 1.1. *An unbiased estimator of the risk difference between the Lasso and the least squares estimator is given by* $\widehat{\Delta R_{10}} = l^2 - 2m$ *where* $l^2 = \|\hat{\beta}_0 - \hat{\beta}_t\|^2_\Sigma$ *and* m *is the codimension of the face of projection of the least squares vector on the Lasso constraint set.*

Note that m equals the number of zero components as estimated by each sequence in the Lasso solution path. Theorem 1.2 provides an unbiased estimator of the risk difference between $\hat{\beta}_\phi$ and the least squares estimator, i.e., it gives an expression for $\widehat{\Delta R_{\phi 0}}$ such that $E_\beta[\widehat{\Delta R_{\phi 0}}] = \Delta R_{\phi 0}$.

Theorem 1.2. *An unbiased estimator of the risk difference between* $\hat{\beta}_\phi$ *and the the least squares estimator is given by* $\widehat{\Delta R_\phi} = \phi^2 l^2 - 2(m-2)\phi - \frac{4\phi}{l^2}\{\hat{\beta}_0^T\Gamma\hat{\beta}_0 - 2\hat{\beta}_0^T\Gamma\hat{\beta}_t + \hat{\beta}_t^T\Gamma\hat{\beta}_t\}$, *where* $\Gamma = \Sigma J$, J *is the Jacobian of the transformation corresponding to the metric projection of the least squares estimator on the Lasso constraint set.*

Define $\widehat{\Delta R} = \widehat{\Delta R_{\phi 0}} - \widehat{\Delta R_{10}}$. This gives an unbiased estimator of $\Delta R = R(\hat{\beta}_\phi, \beta) - R(\hat{\beta}_t, \beta)$. In the spirit of Stein estimation, the following proposition gives a sufficient condition for $\hat{\beta}_\phi$ to be an improved estimator of β over the Lasso.

Theorem 1.3. *Under the conditions of Theorem 1.2,* $\hat{\beta}_\phi$ *dominates the Lasso if* $2 < m < l^2 + 2$ *and* $\phi^2 l^2 - 2(m-2)\phi - \frac{4\phi}{l^2}\{\hat{\beta}_0^T\Gamma\hat{\beta}_0 - 2\hat{\beta}_0^T\Gamma\hat{\beta}_t + \hat{\beta}_t^T\Gamma\hat{\beta}_t\} - l^2 - 2m \leq 0$.

Proof. Define $\widehat{\Delta R} = \widehat{\Delta R_{\phi 0}} - \widehat{\Delta R_{10}}$. It follows that if $\widehat{\Delta R} \leq 0$, then $E[\widehat{\Delta R}] \leq 0$ which further implies $\Delta R_{\phi 0} \leq \Delta R_{10}$. The expression in the proposition follows trivially by plugging in the estimates as obtained in Theorem 1.1 and Theorem 1.2 and noting that $0 < \phi < 1$. □

Note that (1.3) has the following Baranchik form, $\hat{\beta}_\phi = \hat{\beta}_t + (1 - \phi(\cdot))(\hat{\beta}_0 - \hat{\beta}_t)$ where $\phi(\cdot) = r(l^2)(m-2)/l^2$ where $r(l^2)$ is a nondecreasing function of l^2 and $0 < r(l^2) < 1$. Choosing $r(l^2) = \exp(-1/l^2)$ satisfies the requirement. In Section 1.6 we state and derive some properties of treating Lasso as the projection estimator. We derive the divergence of the difference between the least squares estimator and the proposed shrinkage estimator in Proposition 1.2 which is used to prove Theorem 1.2.

1.3 Alternative Shrinkage Estimators

The Lasso produces a sparse model, however, the shrinkage estimator proposed here loses on sparsity but has decision theoretic risk gains over the Lasso estimate. As stated earlier, L_t is a closed convex polytope. Consequently, although the metric projection of the least squares solution onto the polytope is unique, it is not an orthogonal projection. We can get an orthogonal projection if the constraint set is a subspace or a cone instead of a polytope. In this section, we discuss a potential approach of retaining sparsity as obtained by Lasso in addition to conditions of decision theoretic risk domination for a James-Stein type estimator by embedding the Lasso constraint set in a cone.

1.3.1 The Lasso Embedded into a Cone

Recall that the solution set P of a finite system of linear inequalities $A^T X \leq B$ is a convex, possibly vacuous set, and is called a *polyhedron* and a bounded polyhedron is called a poly-

tope. Let $V_1 = (t, 0, \ldots, 0)$, $V_2 = (0, t, \ldots, 0) \ldots, V_p = (0, \ldots, t)$ denote p vectors in $\mathbb{R}^p$. Note that each point V_j has p components. The Lasso constraint set L_t, is the convex hull of the points $V_1, V_2, \ldots, V_p$, referred to as the p-crosspolytope. This is the higher dimensional analogue of the regular octahedron. Consider the following fact from finite dimensional vector spaces, $X = L + L^{\perp}$ where $X \in \mathbb{R}^n$ and $L, L^{\perp}$ are orthogonal subspaces. Such an orthogonal decomposition is not possible if L is a polytope instead of a subspace. However, we can resort to the theory of cones where polar sets play the role which orthogonal complements play in the theory of subspaces. Recall that a set C such that, if $x \in C$ then $\lambda x \in C,\ \forall \lambda \geq 0$ is called a cone (the positive orthant is a simple example). For any arbitrary subset $S \subseteq \mathbb{R}^n$ we define its polar cone as $S^p = \{Y \in \mathbb{R}^n | \langle Y, X \rangle \leq 0, \quad \forall X \in S\}$. Note that S^p is a cone even if S is not. The following Lemma called the Moreau decomposition, can be thought of as an extension of the projection theorem on subspaces to cones.

Lemma 1.1. *Let $C \subseteq \mathbb{R}^n$ be a cone. If $X \in \mathbb{R}^n$ admits an orthogonal decomposition $X = Y + Z$, with $Y \in C$, $Z \in C^p$ and $\langle Y, Z \rangle = 0$, then Y and Z are the projections of X onto C and C^p respectively. Conversely, if $X \in \mathbb{R}^n$ has a projection onto C, then it also has a projection onto C^p and both projections constitute an orthogonal decomposition of X.*

We can rewrite the definition of a polar cone as

$$S^p = \bigcap_{X \in S} \{Y | \langle X, Y \rangle \leq 0\}.$$

In the above equation, the polar cone of any set is being written as an intersection of homogeneous half-spaces. We further have

$$S^{pp} = \bigcap_{Y \in S^p} \{X | \langle Y, X \rangle \leq 0\}.$$

S^{pp} is the intersection of all homogeneous half-spaces which contain S. Note that, if 0 is an

interior point of the set S, then S^p is just the origin and as such, S^{pp} is the whole space, $\mathbb{R}^n$. The above results and the following Lemma can be found in [44].

Lemma 1.2. *S^{pp} is the intersection of all homogeneous half-spaces which contain S. Every polar set S^p is an intersection of half-spaces. Hence $S^{ppp} = S^p$ holds for every subset $S \in \mathbb{R}^n$.*

We associate with an arbitrary set $S \subseteq \mathbb{R}^n$, the cone

$$\mathcal{G}S = \left\{ \begin{pmatrix} X \\ 1 \end{pmatrix} \in \mathbb{R}^{n+1} | X \in S \right\}^{pp} . \tag{1.4}$$

In the other direction, we can define the map

$$\mathcal{D}\tilde{S} = \left\{ X \in \mathbb{R}^n | \begin{pmatrix} X \\ 1 \end{pmatrix} \in \tilde{S} \right\} , \tilde{S} \subseteq \mathbb{R}^{n+1}. \tag{1.5}$$

$\mathcal{G}$ is called the *homogenising operator* and we can think of $\mathcal{D}$ as an "inverse" operator to $\mathcal{G}$.

1.3.2 An Augmented Estimate

In what follows in this section, treat the least squares estimate vector as being an observation in $\mathbb{R}^n$ to keep it distinct from the notation of the polar cone. As defined earlier, let $L_t := \{\beta \in \mathbb{R}^p : \sum_{j=1}^p |\beta_j| \leq t\}$. For the sake of exposition in this section, we assume the tuning parameter t to be fixed. Applying the homogenising operator $\mathcal{G}$, let $C = \mathcal{G}L_t$. Thus C is the smallest cone in $\mathbb{R}^{n+1}$ that contains the n dimensional Lasso constraint set. In other words, we can say that L_t which lies in $\mathbb{R}^n$ has been embedded into the cone C which is in $\mathbb{R}^{n+1}$. To facilitate orthogonal projection of the observed least squares vector onto the cone of constraints, we need to convert the observed least squares in $\mathbb{R}^n$ to an observed vector in $\mathbb{R}^{n+1}$. We thus augment the least squares vector in the following manner.

We know $\hat{\beta}_0 \sim N(\beta, \Sigma^{-1})$. Let $Y \sim N(1, \epsilon)$ be independent of $\hat{\beta}_0$, for some small ϵ. Define the augmented multivariate normal vector as

$$\tilde{X} = \begin{pmatrix} X \\ y \end{pmatrix}$$

where y is a realisation of the normal random variable Y drawn independent of $\hat{\beta}^0$. Thus, $\tilde{X} \sim N(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma})$ where $\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{\beta} \\ 1 \end{pmatrix}$. Also note that

$$C = \left\{ \begin{pmatrix} \boldsymbol{\beta} \\ 1 \end{pmatrix} \in \mathbb{R}^{n+1} | \boldsymbol{\beta} \in L_t \subseteq \mathbb{R}^n \right\}^{pp} . \tag{1.6}$$

Further,

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma^{-1}_{n\times n} & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & \epsilon \end{bmatrix}.$$

Let C^p denote the polar cone of C. Consequently, from Lemma 1.1, we get

$$\widetilde{X} = \widetilde{X_c} + \widetilde{X_{c^p}} \tag{1.7}$$

where $\widetilde{X_c}$ is the projection of $\widetilde{X}$ onto C and $\widetilde{X_{c^p}}$ is the projection of $\widetilde{X}$ onto the polar cone of C and $\langle \widetilde{X_c}, \widetilde{X_{c^p}} \rangle = 0$.

1.3.3 Estimators in the augmented setting

In the augmented problem, we consider $\tilde{\boldsymbol{\beta}}$, the augmented set of parameters (augmented by 1) to lie in the cone $C \in \mathbb{R}^{n+1}$. We could think of the Lasso solution as the restricted maximum likelihood estimate of the original problem, when the parameter space is restricted to lie in a closed convex polytope in $\mathbb{R}^n$. In the augmented problem, the parameter space is

$C \in \mathbb{R}^{n+1}$. For this case, we have $\widetilde{X_c}$ denoting the restricted MLE of $\tilde{\boldsymbol{\beta}}$.

Consider shrinkage estimators based on the restricted maximum likelihood estimate as $(1-\phi)\widetilde{X_c}$. Essentially, we have converted the Lasso problem to an equivalent problem in a dimension higher to be able to take advantage of orthogonal decomposition to cones. Thus, the loss difference in estimating $\tilde{\boldsymbol{\beta}}$ between the shrinkage estimator and the restricted maximum likelihood estimate is

$$\begin{aligned}
\Delta L &= \|(1-\phi)\widetilde{X_c} - \tilde{\boldsymbol{\beta}}\|^2 - \|\widetilde{X_c} - \tilde{\boldsymbol{\beta}}\|^2 \\
&= \|\widetilde{X_{c^p}} + (1-\phi)\widetilde{X_c} - \tilde{\boldsymbol{\beta}}\|^2 - \|\widetilde{X_{c^p}} + \widetilde{X_c} - \tilde{\boldsymbol{\beta}}\|^2 \\
&= \|\widetilde{X_{c^p}} + (1-\phi)\widetilde{X_c} - \tilde{\boldsymbol{\beta}}\|^2 - \|\widetilde{X} - \tilde{\boldsymbol{\beta}}\|^2. \qquad (1.8)
\end{aligned}$$

Therefore, the risk difference is given by

$$\begin{aligned}
\Delta R &= E_\beta[\Delta L] \\
&= E_\beta[\|\widetilde{X_{c^p}} + (1-\phi)\widetilde{X_c} - \tilde{\boldsymbol{\beta}}\|^2 - \|\widetilde{X} - \tilde{\boldsymbol{\beta}}\|^2]
\end{aligned}$$

Writing $\hat{X}^\phi = \widetilde{X_{c^p}} + (1-\phi)\widetilde{X_c}$

$$\begin{aligned}
&= E_\beta[\|\hat{X}^\phi - \tilde{\boldsymbol{\beta}}\|^2] - E_\beta[\|\widetilde{X} - \tilde{\boldsymbol{\beta}}\|^2] \\
&= E_\beta[\|\hat{X}^\phi - \widetilde{X} + \widetilde{X} - \tilde{\boldsymbol{\beta}}\|^2] - E_\beta[\|\widetilde{X} - \tilde{\boldsymbol{\beta}}\|^2] \\
&= E_\beta[\|\widetilde{X} - \hat{X}^\phi\|^2] - 2E_\beta[\langle \widetilde{X} - \tilde{\boldsymbol{\beta}}, \widetilde{X} - \hat{X}^\phi \rangle]
\end{aligned}$$

Applying Lemma 1.3 to the second term in the right hand side above we get,

$$= E_\beta[\phi^2 \|\widetilde{X_c}\|^2 - 2\mathrm{div}\{\phi\widetilde{X_c}\}].$$

Thus, a necessary and sufficient condition for improving over the restricted maximum likelihood estimate is $\phi^2\|\widetilde{X_c}\|^2 - 2\mathrm{div}\{\phi\widetilde{X_c}\} \leq 0$. Notice that the above expression has the exact

same analytical form as (1.28) with $\|\widetilde{X_c}\|^2$ likened to l^2. Note however that the divergence term here is difficult to evaluate owing to a lack of an analytical expression for the point of projection of a vector to a polyhedral cone. Computational evaluation of the projection and the corresponding evaluation of the divergence term is currently work in progress. Recall that $L_t \subset C$. C is the smallest cone in a higher dimension containing the Lasso constraint set. Traversing along the shortest path from $\widetilde{X_c}$ to L_t, one would reach the Lasso constraint set, thus obtaining sparsity along with risk gains over Lasso.

1.3.4 A James-Stein Form

In (1.3), we considered formulation of an estimator in the classical James-Stein framework. In Section 1.3.1, we considered an idea to ensure sparsity in addition to gains in decision theoretic risk where the choice of the shrinkage factor ϕ is clearer. We now look at shrinkage estimators of the Lasso of the following form.

$$\hat{\beta}_{\phi}^{\text{alt}} = (1-\phi)\hat{\beta}_t. \tag{1.9}$$

Define the difference in loss functions due to the estimator in (1.9) and the Lasso as

$$\Delta L_{\text{alt}} = \|\hat{\beta}_{\phi}^{\text{alt}} - \beta\|_{\Sigma}^2 - \|\hat{\beta}_t - \beta\|_{\Sigma}^2. \tag{1.10}$$

The right hand side of (1.10) can be written as

$$\begin{aligned}
\Delta L_{\text{alt}} &= \|\hat{\beta}_{\phi}^{\text{alt}} - \hat{\beta}_t + \hat{\beta}_t - \beta\|_{\Sigma}^2 - \|\hat{\beta}_t - \beta\|_{\Sigma}^2 \\
&= \|\hat{\beta}_{\phi}^{\text{alt}} - \hat{\beta}_t\|_{\Sigma}^2 - 2\langle \hat{\beta}_t - \beta, \hat{\beta}_t - \hat{\beta}_{\phi}^{\text{alt}} \rangle \\
&= \|(1-\phi)\hat{\beta}_t - \hat{\beta}_t\|_{\Sigma}^2 - 2\langle \hat{\beta}_t - \beta, \phi\hat{\beta}_t \rangle.
\end{aligned}$$

which further simplifies to

$$\begin{aligned}\Delta L_{\text{alt}} &= \phi^2\|\hat{\beta}_t\|_\Sigma^2 - 2\langle\hat{\beta}_t - \hat{\beta}_0 + \hat{\beta}_0 - \beta, \phi\hat{\beta}_t\rangle \\ &= \phi^2\|\hat{\beta}_t\|_\Sigma^2 - 2\langle\hat{\beta}_0 - \beta, \phi\hat{\beta}_t\rangle + 2\phi\langle\hat{\beta}_0, \hat{\beta}_t\rangle - 2\phi\|\hat{\beta}_t\|_\Sigma^2 \\ &= (\phi^2 - 2\phi)\|\hat{\beta}_t\|_\Sigma^2 + 2\phi\langle\hat{\beta}_0, \hat{\beta}_t\rangle - 2\langle\hat{\beta}_0 - \beta, \phi\hat{\beta}_t\rangle. \end{aligned} \tag{1.11}$$

Let ΔR_{alt} denote the risk difference between the proposed alternative shrinkage estimator and the Lasso. In other words, ΔR_{alt} denotes the expectation of (1.11). Applying Lemma 1.3 to the last term in (1.11) we get,

$$\Delta R_{\text{alt}} = E_\beta[(\phi^2 - 2\phi)\|\hat{\beta}_t\|_\Sigma^2 + 2\phi\langle\hat{\beta}_0, \hat{\beta}_t\rangle - 2\text{div}\{\phi\hat{\beta}_t\}]. \tag{1.12}$$

where for any p-vector $w = (w_1, w_2, \ldots, w_p)$, $\text{div}(w)$ denotes the divergence of the vector. Observe that all the terms in the content of the expectation operator in (1.12) is independent of the parameter β. In order to compute the expectation of the last term in (1.12), we can use the product rule of divergence as follows. Given a vector F and a scalar valued function ψ, we have

$$\text{div}(\psi F) = \nabla\psi \cdot F + \psi\text{div}(F), \tag{1.13}$$

where $a \cdot b$ denotes the dot product of two vectors a and b. In this case, writing $\phi(l^2) = c/l^2$ for some suitable choice of the constant function c, it can be shown that

$$\begin{aligned}\nabla\phi &= 2\frac{\partial\phi}{\partial l^2}(\hat{\beta}_0^T\Sigma - \hat{\beta}_0^T\Sigma J - \hat{\beta}_t^T\Sigma + \hat{\beta}_t^T\Sigma J) \\ &= 2\frac{\partial\phi}{\partial l^2}(\hat{\beta}_0 - \hat{\beta}_t)^T\Sigma(I - J). \end{aligned} \tag{1.14}$$

where I is the $p \times p$ identity matrix and J is the Jacobian of the transformation corresponding to the metric projection of the least squares vector onto the Lasso constraint set L_t, dervied in Lemma 1.6. From Zou et al. [58], we know that $\text{div}(\hat{\beta}_t) = p - m$ where m is the number

of zeroes as estimated by any solution in the Lasso solution path. We can thus write the divergence term in (1.12) as

$$\text{div}(\phi\hat{\beta}_t) = 2\frac{\partial\phi}{\partial l^2}(\hat{\beta}_0 - \hat{\beta}_t)^T\Sigma(I-J)\cdot\hat{\beta}_t + \phi(p-m) \tag{1.15}$$

Consequently, we get an unbiased estimator of the risk difference between $\hat{\beta}_\phi^{\text{alt}}$ and $\hat{\beta}_t$ as

$$\widehat{\Delta R_{\text{alt}}} = (\phi^2 - 2\phi)\|\hat{\beta}_t\|^2 + 2\phi\langle\hat{\beta}_0, \hat{\beta}_t\rangle - 4\frac{\partial\phi}{\partial l^2}(\hat{\beta}_0 - \hat{\beta}_t)^T\Sigma(I-J)\cdot\hat{\beta}_t - 2\phi(p-m) \tag{1.16}$$

Choosing ϕ such that $\widehat{\Delta R_{\text{alt}}} \leq 0$ almost everywhere and strictly negative on a set of positive measure would give an estimator that would dominate Lasso in this framework. Although such an estimator would have smaller decision theoretic risk compared to the Lasso in addition to retaining sparsity as obtained using the Lasso, a good choice of the shrinkage factor ϕ in this context is not clear, since this functional form of the estimator doesn't fall under the purview of the classical James-Stein framework. This prompts us to propose a shrinkage estimator in the classical framework as given in (1.3).

1.4 Prediction Risk

An important application of the unbiased estimate of risk calculations in the previous section is to model selection. In this section we develop a model and tuning parameter selection methodology. We will apply Efron's ideas [16], where Stein's Unbiased Risk Estimate (SURE) is used to construct estimates of prediction risk. Define, $err_i = (y_i - \hat{\mu}_i)^2$ and $Err_i = E_0(y_i^0 - \hat{\mu}_i)^2$ where E_0 denotes the expectation over $y_i^0 \sim (\mu_i, \sigma^2)$ independent of y with $\hat{\mu}_i$ held fixed. Suppose $\hat{\mu} = \mathbf{m}(\mathbf{y})$ be any rule for estimating μ from $\mathbf{y}$. Using the fact that $E(y_i - \mu_i)^2 = E_0(y_i - \mu_i)^2$ and taking expectations on both sides in the following identity

we get, $(y_i - \mu_i)^2 + (\mu_i - \hat{\mu}_i)^2 = (y_i - \hat{\mu}_i)^2 + 2(\hat{\mu}_i - \mu_i)(y_i - \mu_i)$, so that

$$E(\|y^0 - \hat{\mu}\|)^2 = E(\|y - \hat{\mu}\|)^2 + 2\sum_{i=1}^{n} \text{cov}(\hat{\mu}_i, y_i) \tag{1.17}$$

$$= E(\|y - \hat{\mu}\|)^2 + 2\sigma^2 df(\hat{\mu}).$$

where $df(\hat{\mu}) = \sum_{i=1}^{n} \text{cov}(\hat{\mu}_i, y_i)/\sigma^2$. Note that for a linear estimation rule as in the least squares estimate, the covariance term in (1.17) equals $\sigma^2 H_{ii}$ where H_{ii} denotes the i^{th} diagonal entry of the hat matrix. In the same spirit as the least squares case, the covariance term in (1.17) is also called the *degrees of freedom* of the model. Thus, a covariance penalty is added to the error sum of squares to unbiasedly estimate the prediction risk of the model.

Since $\text{cov}(\hat{\mu}_i, y_i)$ is not an observable statistic, we need some way to estimate it. We use Stein's lemma for the purpose. Suppose $\hat{\mu}_i : \mathbb{R}^n \to \mathbb{R}$ is absolutely continuous in the i^{th} coordinate for $i = 1, \ldots, n$. If $E|\frac{\partial \hat{\mu}_i}{\partial y_i}| < \infty$ for each i, then $\sum_{i=1}^{n} \text{cov}(\hat{\mu}_i, y_i)/\sigma^2 = E(\text{div}\hat{\mu})$ where $\text{div}(\hat{\mu}) = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}$. Hence $\text{div}(\hat{\mu})$ is an unbiased estimator of the degrees of freedom of the model. As such, we can define an Unbiased Estimate of Prediction Risk as follows:

$$PR(\hat{\mu}) = \frac{\|y - \hat{\mu}\|^2}{n} + \frac{2\widehat{df(\hat{\mu})}}{n}\sigma^2. \tag{1.18}$$

Let $\hat{\mu} = X\hat{\beta}_0$ be the ordinary least squares fit and $\hat{\mu}_L = X\hat{\beta}_t$ be the Lasso fit. An unbiased estimate of prediction risk of the ordinary least squares model is given by

$$PR(\hat{\mu}) = \frac{\|y - \hat{\mu}\|^2}{n} + \frac{2\sigma^2}{n}p.$$

An unbiased estimate of the prediction risk of the Lasso model is given by

$$PR(\hat{\mu}_L) = \frac{\|y - \hat{\mu}_L\|^2}{n} + \frac{2\sigma^2}{n}(p - m).$$

where m is the number of zeroes as estimated by Lasso. Let $\hat{\mu}_\phi$ denote the mean estimate using the improved estimator of β, that is let $\hat{\mu}_\phi = X\hat{\beta}_\phi$. Then, we have the following Proposition.

Proposition 1.1. *The prediction risk of $\hat{\mu}_\phi$ is*

$$PR(\hat{\mu}_\phi) = \frac{\|y - \hat{\mu}_\phi\|^2}{n} + \frac{2\widehat{df(\hat{\mu}_\phi)}}{n}\sigma^2. \tag{1.19}$$

where

$$\widehat{df(\hat{\mu}_\phi)} = \mathrm{tr}\left[H - \phi H + \phi\frac{\partial\hat{\mu}_L}{\partial\mathbf{y}} - (\hat{\mu} - \hat{\mu}_L)\frac{\partial\phi}{\partial\mathbf{y}}\right],$$

and

$$\frac{\partial\phi}{\partial\mathbf{y}} = -\frac{2(m-2)}{\sigma^2 l^4}\left[y^T H - \hat{\mu}_L^T - y^T\frac{\partial\hat{\mu}_L}{\partial\mathbf{y}} + \hat{\mu}_L^T\frac{\partial\hat{\mu}_L}{\partial\mathbf{y}}\right].$$

Proof. To develop estimates of the prediction risk, we need to derive expressions for the covariance penalties or equivalently we need to compute the divergence of the model constructed using $\hat{\beta}_\phi$. In what follows, we compute the necessary divergence terms. Since $\hat{\mu}_\phi = X\hat{\beta}_0 - (X\hat{\beta}_0)\phi + (X\hat{\beta}_t)\phi$, taking derivatives on both sides with respect to $\mathbf{y}$ gives,

$$\begin{aligned}\frac{\partial\hat{\mu}_\phi}{\partial\mathbf{y}} &= X\frac{\partial\hat{\beta}_0}{\partial\mathbf{y}} - X\Big(\frac{\partial\hat{\beta}_0}{\partial\mathbf{y}}\Big)\phi - (X\hat{\beta}_0)\Big(\frac{\partial\phi}{\partial\mathbf{y}}\Big) + X\Big(\frac{\partial\hat{\beta}_t}{\partial\mathbf{y}}\Big)\phi + (X\hat{\beta}_t)\Big(\frac{\partial\phi}{\partial\mathbf{y}}\Big)\\ &= H - H\phi + \frac{\partial\hat{\mu}_L}{\partial\mathbf{y}}\phi - (\hat{\mu} - \hat{\mu}_L)\Big(\frac{\partial\phi}{\partial\mathbf{y}}\Big).\end{aligned} \tag{1.20}$$

Note that $\frac{\partial\phi}{\partial\mathbf{y}} = \frac{\partial\phi}{\partial l^2}\frac{\partial l^2}{\partial\mathbf{y}}$. We define l^2 to be the squared weighted distance between the least squares vector and the Lasso, i.e., $l^2 = \|\hat{\beta}_0 - \hat{\beta}_t\|_\Sigma^2 = \hat{\beta}_0^T\Sigma\hat{\beta}_0 - 2\hat{\beta}_0^T\Sigma\hat{\beta}_t + \hat{\beta}_t^T\Sigma\hat{\beta}_t$. Let T_1, T_2, T_3 denote the first, second and third terms in the above expression respectively. Therefore,

$$\frac{\partial l^2}{\partial\mathbf{y}} = \frac{\partial T_1}{\partial\mathbf{y}} - \frac{\partial T_2}{\partial\mathbf{y}} + \frac{\partial T_3}{\partial\mathbf{y}}.$$

Let us look at each term on the right hand side separately.

$$\begin{aligned}\frac{\partial T_1}{\partial \mathbf{y}} &= \hat{\beta}_0{}^T(2\Sigma)\Big(\frac{\partial \hat{\beta}_0}{\partial \mathbf{y}}\Big) \\ &= \frac{2}{\sigma^2} y^T H.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial T_2}{\partial \mathbf{y}} &= 2\hat{\beta}_t{}^T\Sigma\frac{\partial \hat{\beta}_0}{\partial \mathbf{y}} + 2\hat{\beta}_0^T\Sigma\frac{\partial \hat{\beta}_t}{\partial \mathbf{y}} \\ &= \frac{2}{\sigma^2}\hat{\beta}_t^T(X^TX)(X^TX)^{-1}X^T + \frac{2}{\sigma^2}y^TX(X^TX)^{-1}(X^TX)\frac{\partial \hat{\beta}_t}{\partial \mathbf{y}} \\ &= \frac{2}{\sigma^2}\left[\hat{\mu}_L{}^T + y^T\frac{\partial \hat{\mu}_L}{\partial \mathbf{y}}\right].\end{aligned}$$

It can be shown that

$$\frac{\partial T_3}{\partial \mathbf{y}} = 2\hat{\beta}_t^T\Sigma\Big(\frac{\partial \hat{\beta}_t}{\partial \mathbf{y}}\Big).$$

Combining the three terms it follows that,

$$\frac{\partial l^2}{\partial \mathbf{y}} = \frac{2}{\sigma^2}\left[y^TH - \hat{\mu}_L^T - y^T\frac{\partial \hat{\mu}_L}{\partial \mathbf{y}} + \hat{\mu}_L^T\frac{\partial \hat{\mu}_L}{\partial \mathbf{y}}\right].$$

Further note that $\partial\phi/\partial l^2 = -(m-2)/l^4$. Thus, plugging these into (1.20), we get

$$\frac{\partial \hat{\mu}_\phi}{\partial \mathbf{y}} = H - \phi H + \phi\frac{\partial \hat{\mu}_L}{\partial \mathbf{y}} - (\hat{\mu} - \hat{\mu}_L)\frac{\partial \phi}{\partial \mathbf{y}},$$

where

$$\frac{\partial \phi}{\partial \mathbf{y}} = -\frac{2(m-2)}{\sigma^2 l^4}\left[y^TH - \hat{\mu}_L^T - y^T\frac{\partial \hat{\mu}_L}{\partial \mathbf{y}} + \hat{\mu}_L^T\frac{\partial \hat{\mu}_L}{\partial \mathbf{y}}\right].$$

Taking trace on both sides of the above equation gives an expression for the degrees of freedom. □

1.5 Simulation

We perform a simulation study to check the effectiveness of the proposed estimator as compared to both the Lasso and the OLS. We use Stein's unbiased estimate of risk (prediction risk) as the criterion for choosing the estimator and report the corresponding estimates of decision theoretic parameter risk for the proposed estimate in (1.3) (a James-Stein type estimator) and the Lasso estimator. We also report the corresponding data-based shrinkage factors. Note that we are operating under the $p < n$ case.

Here, we present three different simulation settings. The data in Table 1.1 corresponds to the case where we set $n = 200, p = 100, \sigma^2 = 1$. The data in Table 1.3 corresponds to the case where we set $n = 200, p = 100, \sigma^2 = 9000$ and the data in Table 1.4 corresponds to the case where we set $n = 2000, p = 100, \sigma^2 = 9000$. and the All the three tables have 9 columns. The first column, denoted by A gives the size of the active set for the model. In other words, it shows the varying levels of sparsity under which the simulation study is conducted. The active set decreases from 90 to 10 where 90 indicates that the true number of zeroes in the model are 10. Similarly, $A = 10$ indicates that the true number of zeroes in the linear model is 90. The second column denoted by 'RiskDiff' gives the estimates of the difference in decision theoretic parameter risk between the James-Stein type estimator and the Lasso. A more negative value in this column corresponds to more risk gains by doing James-Stein estimation. The third, fourth and fifth columns present the estimates of Prediction Risk for (1.3), the Lasso and the least squares model respectively. The sixth column displays the shrinkage factor for the corresponding James-Stein estimator. The last three columns give the 'True Loss' for (1.3), the Lasso and the least squares methods, respectively. The true loss is the weighted l_2 norm between the estimate and the true parameter value, weighted

by the design matrix. The simulation methodology is outlined in the next paragraph.

Generate a p dimensional β^* vector with A nonzero components in β^*. Generate an $n \times p$ matrix X, from a multivariate normal distribution such that the correlation between X_i and X_j is $\rho = 0.5^{|i-j|}$. Setting $y_i = \sum_{j=1}^{p} x_{ij}\beta_j^* + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2{}_\epsilon)$, we generate the response y. Performing the LARS algorithm on the generated dataset, we get the entire LASSO solution path and pick the optimal Lasso solution for which the estimate of Stein's unbiased risk estimate (SURE) is the smallest. For each step in the path, compute $l^2 = \|\hat{\beta}_0 - \hat{\beta}_t\|_\Sigma^2$ and m denoting the number of zero components as estimated by each solution in the Lasso solution path. Further, set the shrinkage factor as $\phi(l) = (m-2)/l^2$. For each solution in the path, compute the corresponding shrinkage estimate of β as $\hat{\beta}_\phi = \hat{\beta}_t + (1-\phi)(\hat{\beta}_0 - \hat{\beta}_t)$, where $\hat{\beta}_t$ is a solution in the solution path and $\hat{\beta}_0$ is the least squares solution. We then compute $\widehat{\Delta R_{10}} = l^2 - 2m$ and $\widehat{\Delta R_{\phi 0}} = \phi^2 l^2 - 2(m-2)\phi - \frac{4(m-2)}{l^4}\{\hat{\beta}_0{}^T \Gamma \hat{\beta}_0 - 2\hat{\beta}_0{}^T \Gamma \hat{\beta}_t + \hat{\beta}_t{}^T \Gamma \hat{\beta}_t\}$ corresponding to each Lasso solution in the path. Compute estimates of Prediction Risk for the model obtained by using the shrinkage estimates of β from analytical expressions as obtained in Section 1.4 and pick the Shrinkage solution corresponding to the smallest estimate of SURE. Finally, report the corresponding estimates of the risk difference of the parameters as $\widehat{\Delta R} = \widehat{\Delta R_\phi} - \widehat{\Delta R_L}$.

From the data in Tables 1.1, 1.3 and 1.4 we can see that the amount of risk gains is directly related to the active set in the model. Comparing the three tables, we can also observe that a larger value of n for a fixed value of p yields more gains in risk for similar values of the active set. For instance, when $n = 2000$, $A = 90$ and $\sigma^2 = 9000$, we see risk gains of the shrinkage estimator over Lasso to be around 15 times more compared to when $n = 200$ for the same

values of A and σ^2. The risk gains become less as the cardinality of the active set decreases. It is further noticeable comparing the results in 1.1 and 1.3 that the proposed estimator is invariant to the error variance in the model. The pattern of risk gains is similar when we compare the results of these two tables. We also observe that estimates of prediction risks for the shrinkage model and the Lasso model are very similar across every simulation setting. As expected, both methods consistently have smaller prediction risks compared to the least squares estimator.

In Figures 1.1, 1.2 and 1.3, we give the plots of the risk difference between the shrinkage model and the Lasso model as a function of the shrinkage factor. It can be seen that shrinkage factors closer to 0 give more risk gains of the proposed estimator over the Lasso. However, when shrinkage factor is closer to 1, the risk gains are quite low and on some occasions, Lasso does marginally better than the proposed estimator. Shrinkage factor closer to 0 corresponds to a higher cardinality of the active set in the model and shrinkage factors closer to 1 correspond to smaller active sets in the model. This supports the theory in the sense that $\phi = 1$ reduces to the Lasso, which acts like a boundary case for the proposed estimator.

We then conduct the analysis on the diabetes dataset. The data in Tables 1.5, 1.6 and 1.7 summarise the results obtained by conducting the analysis on the diabetes data. The dataset has $p = 64$ predictors on $n = 442$ observations. Each row of the table gives the number of predictors used in the model which is increased from $p = 10$ to $p = 64$, the active set A, giving the number of nonzero predictors in the model as obtained by doing the Lasso using the LARS algorithm and the corresponding risk difference between (1.3) and the Lasso along with the respective prediction risk estimates of improved estimate and the Lasso. Figure 1.4

shows the optimal proposed estimate, the Lasso and the ordinary least squares estimator as a function of the number of predictors in the dataset. Optimality criterion is adjudged by the model for which the estimate of prediction risk is the smallest. Figure 1.5 depicts the decision theoretic risk difference between the optimal proposed model and the optimal Lasso fit as a function of the number of predictors used in the model. We observe substantial risk gains through the entire range of the diabetes data set when $A/p < \delta$ for some $0 < \delta < 1$. It can also be seen that our proposed estimator also has smaller Prediction risks compared to the Lasso for most of the range of the diabetes data. A theoretical understanding of the threshold δ is work in progress.

1.6 Proofs of Section 1.2.1 Results

The following lemma from Fourdrinier et al. [19] is used repeatedly in this chapter and hence we state it here for the sake of completeness.

Lemma 1.3. *If $X \sim N_p(\theta, \Sigma)$, where Σ is a positive definite symmetric matrix. Let $L(X, \theta) = \|X - \theta\|^2_\Sigma = (X - \theta)^T \Sigma^{-1} (X - \theta)$ denote the general quadratic loss function and $\langle \cdot, \cdot \rangle$ denote the inner product induced by this norm. Then, for a weakly differentiable function $g : \mathbb{R}^p \to \mathbb{R}^p$, we have, $E_\theta[\langle X - \theta, g(X) \rangle] = E_\theta[\text{div} g(X)]$, where $\text{div} g(X)$ denotes the divergence of $g(X)$.*

It is a well known classical result, see for example Deutsch [11], that if K is a non-empty closed and convex subset of $\mathbb{R}^p$, then the metric projection of any point $x \in \mathbb{R}^p$ onto K is unique. In a linear model setting, $Y = X\beta + \epsilon$, the least squares solution, $\hat{\beta}_0$ is a random vector in $\mathbb{R}^p$. The Lasso constraint set, given by $L_t := \{\beta \in \mathbb{R}^p : \sum_{j=1}^p |\beta_j| \leq t\}$, is a closed and convex subset of $\mathbb{R}^p$. Let $\hat{\beta}_{L_t}$ denote the unique metric projection of the least squares

estimator onto L_t. We call $\hat{\beta}_{L_t}$ as the projection estimator and is unique for any fixed value of the tuning parameter t. Define the inner product $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^p$ by $\langle w_1, w_2 \rangle = w_1' V w_2$ for $w_1, w_2 \in \mathbb{R}^p$ where $V = (X^T X)$. Then, we note from Kato [27] that the Lasso solution can be treated as the projection of the least squares estimator onto the cross-polytope. Thus, in what follows, the Lasso and the projection estimator would be used interchangeably, however, we shall call the LARS output as the Lasso estimate.

We know that $\hat{\beta}_0 \sim N(\beta, \sigma^2 (X^T X)^{-1})$. Let $\Sigma^{-1} = \sigma^2 (X^T X)^{-1}$. Following the same notation convention as in Zou et al. [58], let $\hat{\mu}_\lambda$ be the Lasso fit for any value of λ as obtained using the LARS algorithm. Suppose $\mathbf{M}$ is any matrix with p columns. Let $\mathcal{S}$ be a subset of the indices $\{1, 2, \ldots, p\}$. Define $\beta_{\mathcal{S}} = (\ldots \beta_j \ldots)_{j \in \mathcal{S}}$ for any vector β of length p. Let $\text{sgn}(\cdot)$ be the sign function: $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$ and $\text{sgn}(x) = 0$ if $x = 0$. Let $\mathcal{A}$ denote the active set of β, i.e., the set of β for which $\text{sgn}(\beta) \neq 0$ where $\text{sgn}(\beta)$ is the sign vector of β given by $\text{sgn}(\beta)_j = \text{sgn}(\beta_j)$. Let $\mathcal{B}_m$ denote the active set at the transition points, λ_m. Let the corresponding submatrix of X be denoted by $X_{\mathcal{B}_m}$. The entire Lasso solution path is obtained by performing the LARS algorithm as given in Efron et al. [17]. For a given response vector y, there is a finite sequence of λ's called the transition points,

$$\lambda_0 > \lambda_1 > \lambda_2 > \ldots > \lambda_K$$

such that for all $\lambda > \lambda_0$, $\hat{\beta}_t(\lambda) = 0$. We state the following Lemma from [58] for the sake of completeness.

Lemma 1.4. Zou et al.(2007) *Consider the transition points λ_m and λ_{m+1}, $\lambda_{m+1} \geq 0$. $\mathcal{B}_m$ is the active set in $(\lambda_{m+1}, \lambda_m)$. Suppose i_{add} is an index added into $\mathcal{B}_m$ at λ_m and its index in $\mathcal{B}_m$ is i^*, that is, $i_{add} = (\mathcal{B}_m)_{i^*}$. Denote by $(a)_k$, the k^{th} element of the vector a. We can*

express the transition point λ_m *as*

$$\lambda_m = \frac{2((X_{\mathcal{B}_m} X_{\mathcal{B}_m})^{-1} X_{\mathcal{B}_m}^T y)_{i^*}}{((X_{\mathcal{B}_m} X_{\mathcal{B}_m})^{-1} \mathrm{sgn}_m)_{i^*}}.$$

Moreover, if j_{drop} *is a dropped index at* λ_{m+1} *and* $j_{drop} = (\mathcal{B}_m)_{j^*}$*, then* λ_{m+1} *can be written as*

$$\lambda_{m+1} = \frac{2((X_{\mathcal{B}_m} X_{\mathcal{B}_m})^{-1} X_{\mathcal{B}_m}^T y)_{j^*}}{((X_{\mathcal{B}_m} X_{\mathcal{B}_m})^{-1} \mathrm{sgn}_m)_{j^*}}.$$

Denoting $\partial/\partial y$ as the partial derivative with respect to the vector $y \in \mathbb{R}^n$ and $\partial/\partial\hat{\beta}_0$ as the partial derivative with respect to the least squares estimates $\hat{\beta}_0 \in \mathbb{R}^p$ we have the following Lemma.

Lemma 1.5. *Let* $\hat{\mu}_L = X\hat{\beta}_t$ *be the Lasso fit for any value of* t *or equivalently for any* λ *. Then,* $\mathrm{div}(\hat{\mu}_L) = \mathrm{tr}[\partial\hat{\beta}_t/\partial\hat{\beta}_0] = \mathrm{tr}J$*.*

Proof. The result follows from the following divergence calculation.

$$\mathrm{div}(\hat{\mu}_L) = \mathrm{tr}\left[\frac{\partial\hat{\mu}_L}{\partial y}\right] = \mathrm{tr}\left[X\frac{\partial\hat{\beta}_t}{\partial\hat{\beta}_0}\frac{\partial\hat{\beta}_0}{\partial y}\right] = \mathrm{tr}\left[\frac{\partial\hat{\beta}_t}{\partial\hat{\beta}_0}\right].$$

□

We can define the Jacobian of the projection of the least squares estimator on the Lasso polytope. The first part of the following Lemma gives the Jacobian of the transformation for the case when λ is a non-transition point and the second part gives the Jacobian when λ is a transition point.

Lemma 1.6. *(i) The Jacobian of the projection of the ordinary least squares vector onto the closed convex polytope when* $\lambda \in (\lambda_{m+1}, \lambda_m)$*, i.e., when* λ *is a non-transition point is given by*

$$J = \frac{\partial\hat{\beta}_t}{\partial\hat{\beta}_0} = (X^T X)^{-1} X^T H_\lambda(y) X$$

where $H_\lambda(y)$ is the projection matrix on the subspace of the nonzero coordinates of the Lasso solution, i.e., on the subspace spanned by the vectors in the active set, $\mathcal{A}$.

(ii) The Jacobian of the projection of the ordinary least squares vector onto the closed convex polytope at the transition points, λ_m is given by

$$J = \frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0} = (X^T X)^{-1} X^T \Phi_m X$$

where

$$\Phi_m = H_{\mathcal{B}_m} - \frac{1}{2} X_{\mathcal{B}_m} (X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1} \text{sgn}_m \vartheta(\mathcal{B}_m, i^*) X_{\mathcal{B}_m}^T$$

and

$$\vartheta(\mathcal{B}_m, i^*) = 2 \left\{ \frac{(X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1}[i^*, \cdot]}{(X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1}[i^*, \cdot] \text{sgn}_m} \right\}.$$

Proof. We know from Zou et al. [58] that when $\lambda \in (\lambda_{m+1}, \lambda_m)$, a nontransition point, the active set $\mathcal{A}(\lambda)$ and the sign vector $\text{sgn}(\lambda)$ are locally constant. Further, $\hat{\mu}_\lambda(y)$ is uniformly Lipschitz in y. As such we get,

$$\frac{\partial \hat{\mu}_\lambda(y)}{\partial y} = H_\lambda(y). \tag{1.21}$$

Thus,

$$\begin{aligned} H_\lambda(y) &= X \frac{\partial \hat{\beta}_t}{\partial y} \\ &= X \left(\frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0} \right) \left(\frac{\partial \hat{\beta}_0}{\partial y} \right) \\ &= X \left(\frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0} \right) (X^T X)^{-1} X^T \\ X^T H_\lambda(y) &= (X^T X) \left(\frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0} \right) (X^T X)^{-1} X^T. \end{aligned}$$

From the above expression, part (i) of the Lemma follows as

$$\frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0} = (X^T X)^{-1} X^T H_\lambda(y) X. \tag{1.22}$$

To demonstate part (ii) of the Lemma, note that the Lasso fit at the transition points is given by

$$\mu_m(y) = H_{\mathcal{B}_m} y - \frac{1}{2} X_{\mathcal{B}_m} (X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1} \text{sgn}_m \lambda_m.$$

Note that at the transition points, sgn_m is strictly nonzero. Thus, differentiating both sides with respect to y, we get,

$$\begin{aligned}
\frac{\partial \mu_m(y)}{\partial y} &= H_{\mathcal{B}_m} - \frac{1}{2} X_{\mathcal{B}_m} (X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1} \text{sgn}_m \frac{\partial \lambda_m}{\partial y} \\
\frac{\partial \mathcal{B}_m}{\partial y} &= (X^T X)^{-1} X^T \{ H_{\mathcal{B}_m} - \frac{1}{2} X_{\mathcal{B}_m} (X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1} \text{sgn}_m \frac{\partial \lambda_m}{\partial y} \} X.
\end{aligned}$$

From Lemma 1.4, we can get that,

$$\begin{aligned}
\frac{\partial \lambda_m}{\partial y} &= 2 \left\{ \frac{(X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1}[i^*, \cdot]}{(X_{\mathcal{B}_m}^T X_{\mathcal{B}_m})^{-1}[i^*, \cdot] \text{sgn}_m} \right\} X_{\mathcal{B}_m}^T \\
&= \vartheta(\mathcal{B}_m, i^*) X_{\mathcal{B}_m}^T.
\end{aligned}$$

□

Proposition 1.2. *Assume $\Sigma = (X^T X)/\sigma^2$ is completely specified and the Jacobian J, is as in Lemma 1.6, then*

$$\text{div}(\hat{\beta}_0 - \hat{\beta}_\phi) = (m-2)\phi + \frac{2\phi}{l^2} \left[\hat{\beta}_0^T \Gamma \hat{\beta}_0 - 2\hat{\beta}_0^T \Gamma \hat{\beta}_t + \hat{\beta}_t^T \Gamma \hat{\beta}_t \right].$$

where $\Gamma = \Sigma J$.

Proof. Note that $\hat{\beta}_\phi = \hat{\beta}_t + (1-\phi)(\hat{\beta}_0 - \hat{\beta}_t)$ where $\phi = (m-2)/l^2$. Then,

$$\begin{aligned}
\operatorname{div}(\hat{\beta}_0 - \hat{\beta}_\phi) &= \operatorname{div}\{\phi(\hat{\beta}_0 - \hat{\beta}_t)\} \\
&= \sum_{j=1}^{p} \frac{\partial}{\partial \hat{\beta}_{0j}} \{\phi(\hat{\beta}_{0j} - \hat{\beta}_{tj})\} \\
&= \sum_{j=1}^{p} \{\phi \frac{\partial}{\partial \hat{\beta}_{0j}} (\hat{\beta}_{0j} - \hat{\beta}_{tj}) + (\hat{\beta}_{0j} - \hat{\beta}_{tj}) \frac{\partial \phi}{\partial \hat{\beta}_{0j}}\} \\
&= m\phi + \sum_{j=1}^{p} (\hat{\beta}_{0j} - \hat{\beta}_{tj}) \frac{\partial \phi}{\partial \hat{\beta}_{0j}} \\
&= m\phi - \frac{(m-2)}{l^4} \sum_{j=1}^{p} (\hat{\beta}_{0j} - \hat{\beta}_{tj}) \frac{\partial l^2}{\partial \hat{\beta}_{0j}}.
\end{aligned}$$

Note that l^2 is a scalar and $\hat{\beta}_0$ is a $(p \times 1)$ vector and so $\partial l^2 / \partial \hat{\beta}_0$ is a $(1 \times p)$ vector. We can write the above expression as

$$\operatorname{div}\{\phi(\hat{\beta}_0 - \hat{\beta}_t)\} = m\phi - \frac{\phi}{l^2} \left\{ \left(\frac{\partial l^2}{\partial \hat{\beta}_0} \right) \cdot (\hat{\beta}_0 - \hat{\beta}_t) \right\}. \tag{1.23}$$

Note also that $l^2 = \|\hat{\beta}_0\|_\Sigma^2 - 2\langle \hat{\beta}_0, \hat{\beta}_t \rangle + \|\hat{\beta}_t\|_\Sigma^2 = \hat{\beta}_0^T \Sigma \hat{\beta}_0 - 2\hat{\beta}_0^T \Sigma \hat{\beta}_t + \hat{\beta}_t^T \Sigma \hat{\beta}_t$. Let T_1, T_2, T_3 denote the first, second and third terms in the above expression respectively. Note that T_1 and T_3 are quadratic forms and T_2 is a bilinear form, thus it follows that

$$\frac{\partial l^2}{\partial \hat{\beta}_0} = \frac{\partial T_1}{\partial \hat{\beta}_0} - \frac{\partial T_2}{\partial \hat{\beta}_0} + \frac{\partial T_3}{\partial \hat{\beta}_0}.$$

Consequently we obtain,

$$\frac{\partial l^2}{\partial \hat{\beta}_0} \cdot (\hat{\beta}_0 - \hat{\beta}_t) = \left(\frac{\partial T_1}{\partial \hat{\beta}_0} - \frac{\partial T_2}{\partial \hat{\beta}_0} + \frac{\partial T_3}{\partial \hat{\beta}_0} \right) \cdot (\hat{\beta}_0 - \hat{\beta}_t).$$

Now,

$$\frac{\partial T_1}{\partial \hat{\beta}_0} \cdot (\hat{\beta}_0 - \hat{\beta}_t) = 2\|\hat{\beta}_0\|_\Sigma^2 - 2\langle \hat{\beta}_0, \hat{\beta}_t \rangle.$$

$$\begin{aligned}\frac{\partial T_2}{\partial \hat{\beta}_0} \cdot (\hat{\beta}_0 - \hat{\beta}_t) &= (2\hat{\beta}_t^T \Sigma + 2\hat{\beta}_0^T \Sigma \frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0}) \cdot (\hat{\beta}_0 - \hat{\beta}_t) \\ &= 2\langle \hat{\beta}_0, \hat{\beta}_t \rangle - 2\|\hat{\beta}_t\|_\Sigma^2 + 2\hat{\beta}_0^T \Gamma \hat{\beta}_0 - 2\hat{\beta}_0^T \Gamma \hat{\beta}_t.\end{aligned}$$

$$\frac{\partial T_3}{\partial \hat{\beta}_0} \cdot (\hat{\beta}_0 - \hat{\beta}_t) = 2\hat{\beta}_t^T \Gamma \hat{\beta}_0 - 2\hat{\beta}_t^T \Gamma \hat{\beta}_t.$$

where $\Gamma = \Sigma(\partial \hat{\beta}_t / \partial \hat{\beta}_0)$. Combining the above equations we get,

$$\left(\frac{\partial l^2}{\partial \hat{\beta}_0}\right) \cdot (\hat{\beta}_0 - \hat{\beta}_t) = 2[l^2 - \hat{\beta}_0^T \Gamma \hat{\beta}_0 - \hat{\beta}_t^T \Gamma \hat{\beta}_t + 2\hat{\beta}_0^T \Gamma \hat{\beta}_t]. \tag{1.24}$$

Substituting (1.24) in the second term of (1.23) we get,

$$\begin{aligned}\text{div}\{\phi(\hat{\beta}_0 - \hat{\beta}_t)\} &= m\phi - \frac{2\phi}{l^2}\Big[l^2 - \hat{\beta}_0^T \Gamma \hat{\beta}_0 - \hat{\beta}_t^T \Gamma \hat{\beta}_t + 2\hat{\beta}_0^T \Gamma \hat{\beta}_t\Big] \\ &= (m-2)\phi + \frac{2\phi}{l^2}\Big[\hat{\beta}_0^T \Gamma \hat{\beta}_0 - 2\hat{\beta}_0^T \Gamma \hat{\beta}_t + \hat{\beta}_t^T \Gamma \hat{\beta}_t\Big].\end{aligned}$$

□

We now prove Theorems 1.1 and 1.2.

PROOF OF THEOREM 1.1: Let

$$\begin{aligned}\Delta R_{10} &= E[\|\hat{\beta}_t - \beta\|_\Sigma^2 - \|\hat{\beta}_0 - \beta\|_\Sigma^2] \\ &= E[\|\hat{\beta}_t - \hat{\beta}_0 + \hat{\beta}_0 - \beta\|_\Sigma^2 - \|\hat{\beta}_0 - \beta\|_\Sigma^2] \\ &= E[\|\hat{\beta}_0 - \beta\|_\Sigma^2 + \|\hat{\beta}_0 - \hat{\beta}_t\|_\Sigma^2 - 2\langle \hat{\beta}_0 - \beta, \hat{\beta}_0 - \hat{\beta}_t \rangle \\ &\quad - \|\hat{\beta}_0 - \beta\|_\Sigma^2] \\ &= E[\|\hat{\beta}_0 - \hat{\beta}_t\|_\Sigma^2 - 2\langle \hat{\beta}_0 - \beta, \hat{\beta}_0 - \hat{\beta}_t \rangle] \\ &= E[l^2 - 2\langle \hat{\beta}_0 - \beta, \hat{\beta}_0 - \hat{\beta}_t \rangle].\end{aligned}$$

Letting $g(\hat{\beta}_0) = \hat{\beta}_0 - \hat{\beta}_t$ and using Lemma 1.3 we get,

$$= E[l^2 - 2\text{div}(\hat{\beta}_0 - \hat{\beta}_t)] \tag{1.25}$$

where $l^2 = \|\hat{\beta}_0 - \hat{\beta}_t\|_\Sigma^2$. We next need an expression for the divergence term in the above equation. Note that,

$$\begin{aligned}\text{div}(\hat{\beta}_0 - \hat{\beta}_t) &= \sum_{j=1}^{p} \frac{\partial}{\partial \hat{\beta}_{0j}}(\hat{\beta}_{0j} - \hat{\beta}_{tj}) \\ &= p - \text{tr}\left(\frac{\partial \hat{\beta}_t}{\partial \hat{\beta}_0}\right).\end{aligned} \tag{1.26}$$

From Zou et al. [58], $\text{tr}(\partial\hat{\beta}_t/\partial\hat{\beta}_0) = p - m$. We thus, get an expression for an unbiased estimator of risk difference between the Lasso and least squares estimate as

$$\widehat{\Delta R_{10}} = l^2 - 2m. \tag{1.27}$$

□

PROOF OF THEOREM 1.2:

$$\begin{aligned}\Delta R_{\phi 0} &= E_\beta[\|\hat{\beta}_\phi - \beta\|_\Sigma^2] - E_\beta[\|\hat{\beta}_0 - \beta\|_\Sigma^2] \\ &= E_\beta[\|\hat{\beta}_\phi - \hat{\beta}_0 + \hat{\beta}_0 - \beta\|_\Sigma^2 - \|\hat{\beta}_0 - \beta\|_\Sigma^2] \\ &= E_\beta[\|\hat{\beta}_0 - \beta\|_\Sigma^2 + \|\hat{\beta}_0 - \hat{\beta}_\phi\|_\Sigma^2 - 2\langle\hat{\beta}_0 - \beta, \hat{\beta}_0 - \hat{\beta}_\phi\rangle \\ &\quad - \|\hat{\beta}_0 - \beta\|_\Sigma^2] \\ &= E_\beta[\|\hat{\beta}_0 - \hat{\beta}_\phi\|_\Sigma^2 - 2\langle\hat{\beta}_0 - \beta, \hat{\beta}_0 - \hat{\beta}_\phi\rangle].\end{aligned}$$

Letting $g(\hat{\beta}_0) = \hat{\beta}_0 - \hat{\beta}_\phi$ and using Lemma 1.3 we get,

$$\begin{aligned}
&= E_\beta[\|\hat{\beta}_0 - \hat{\beta}_\phi\|_\Sigma^2 - 2\text{div}(\hat{\beta}_0 - \hat{\beta}_\phi)] \\
&= E_\beta[\|\phi(\hat{\beta}_0 - \hat{\beta}_t)\|_\Sigma^2 - 2\text{div}(\hat{\beta}_0 - \hat{\beta}_\phi)] \\
&= E_\beta[\phi^2 l^2 - 2\text{div}(\hat{\beta}_0 - \hat{\beta}_\phi)] \\
&= E_\beta[\phi^2 l^2 - 2\text{div}\{\phi(\hat{\beta}_0 - \hat{\beta}_t)\}]. \qquad (1.28)
\end{aligned}$$

From Propostion 1.2, we get an expression for the divergence term. We thus have,

$$\Delta R_{\phi 0} = E\left[\phi^2 l^2 - 2(m-2)\phi - \frac{4\phi}{l^2}\{\hat{\beta}_0^T \Gamma \hat{\beta}_0 - 2\hat{\beta}_0^T \Gamma \hat{\beta}_t + \hat{\beta}_t^T \Gamma \hat{\beta}_t\}\right]. \qquad (1.29)$$

□

1.7 Discussion

We have presented an approach to understand Stein estimation in the penalised regression framework establishing analytical results for risk estimates of the Lasso and the corresponding Stein type estimators obtained by treating the Lasso as restricted maximum likelihood estimate. In addition, we have estimates of the degrees of freedom of the shrinkage model which in turn provides unbiased estimates of prediction risk for optimal tuning parameter selection.

A possible next step of the above approach would be to evaluate the above risk expressions by retaining the sparsity that Lasso produces. Yet another extension would be understand how the shrinkage model works for the $p > n$ case. Further, we need to establish asymptotic properties of the shrinkage estimator focussing in particular to evaluate asymptotic risk of the Stein-type estimator and also deriving the limit distribution of the proposed estimator.

In addition, deriving the distribution of the squared length of projection of the least squares solution on the cross-polytope can help understand the risk properties of the Lasso better.

1.8 Tables and Figures

Table 1.1 – JS Estimate versus Lasso for $n = 200, p = 100, {\sigma^2}_\epsilon = 1$

A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage	LossJS	LossLasso	LossOLS
90	-45.885	1.485	1.482	1.489	0.1865	99.18	105.11	99.17
80	-7.965	1.481	1.473	1.498	0.4515	100.31	100.09	100.36
70	-4.069	1.455	1.446	1.491	0.6083	98.63	95.51	99.09
60	-0.715	1.433	1.420	1.493	0.7719	96.37	91.86	100.52
50	-0.013	1.407	1.394	1.497	0.7623	90.03	87.27	100.36
40	0.565	1.363	1.348	1.491	0.7940	80.23	78.66	99.43
30	1.591	1.318	1.302	1.494	0.8476	69.99	69.22	100.52
20	1.758	1.252	1.234	1.490	0.8687	58.51	57.66	100.34
10	2.364	1.170	1.150	1.495	0.8982	40.81	38.55	100.21

Table 1.2 – JS Estimate versus Lasso for $n = 200, p = 100, {\sigma^2}_\epsilon = 100$

A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage	LossJS	LossLasso	LossOLS
90	-81.826	148.27	147.71	149.70	0.3426	98.97	101.50	100.30
80	-12.193	146.78	145.89	149.52	0.5282	97.55	98.26	100.98
70	-5.622	144.68	143.88	149.38	0.6286	93.80	94.22	100.43
60	-0.8933	141.88	140.68	149.38	0.7361	88.97	89.11	99.88
50	-0.2796	139.57	138.25	150.02	0.7653	82.69	82.57	99.58
40	0.6376	134.95	133.34	149.34	0.8086	75.78	75.48	99.60
30	1.0263	130.90	129.35	149.93	0.8406	67.97	67.16	100.97
20	1.6463	124.04	122.09	149.22	0.8700	55.42	54.29	99.89
10	2.1238	116.77	114.75	149.97	0.8948	37.97	35.92	100.95

Table 1.3 – JS Estimate versus Lasso for $n = 200, p = 100, {\sigma^2}_\epsilon = 9000$

A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage	LossJS	LossLasso	LossOLS
90	-16.213	11691.05	11683.32	13441.48	0.6959	68.05	70.75	100.68
80	-11.186	11530.18	11526.61	13442.60	0.7276	64.56	67.22	100.27
70	-9.516	11384.45	11369.39	13467.62	0.7430	59.51	62.23	99.66
60	-6.569	11147.69	11105.46	13457.92	0.7704	55.40	57.44	99.41
50	-4.567	10925.98	10885.12	13451.05	0.7963	49.38	51.39	99.65
40	-1.586	10609.42	10545.22	13422.45	0.8255	44.28	46.25	100.41
30	-0.074	10283.63	10184.81	13429.07	0.8526	37.18	38.81	100.99
20	1.049	10079.60	9957.54	13503.79	0.8763	28.05	29.01	100.96
10	4.018	9809.09	9419.89	13426.02	0.8958	16.86	16.98	100.39

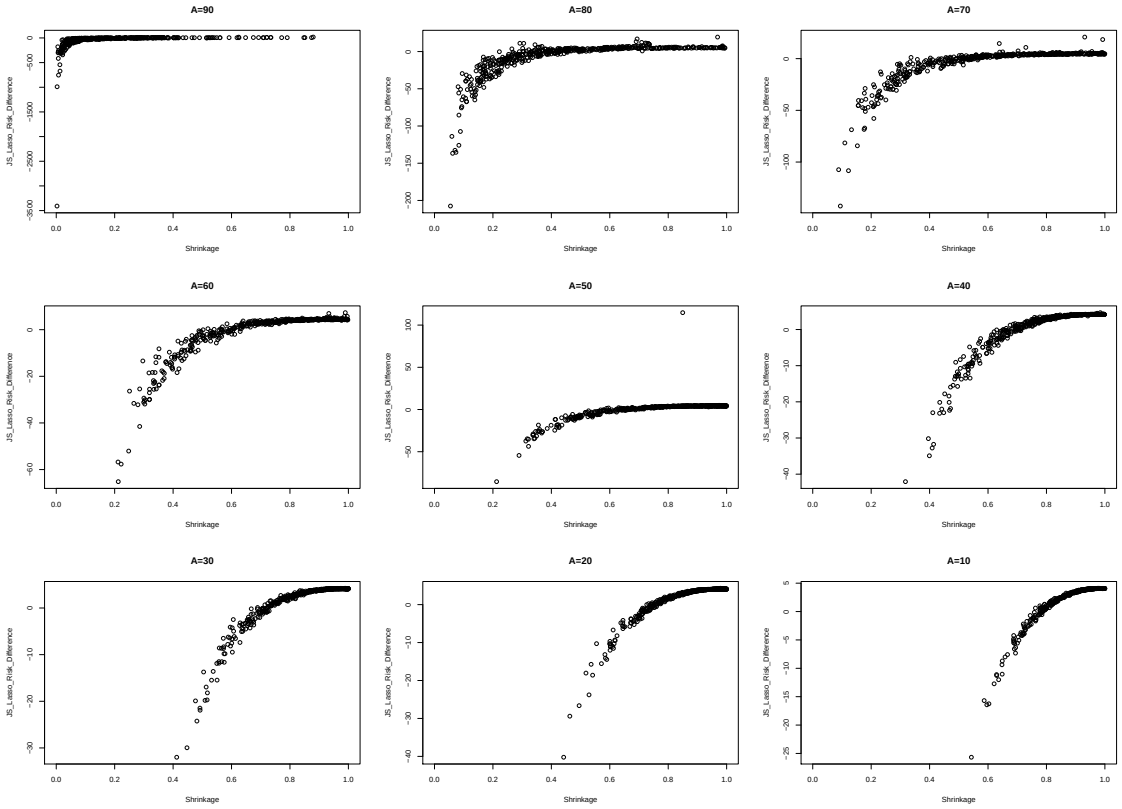

Figure 1.1 – $n = 200, p = 100, \sigma^2 = 1$. Shrinkage Factor vs JS Estimate Risk Difference

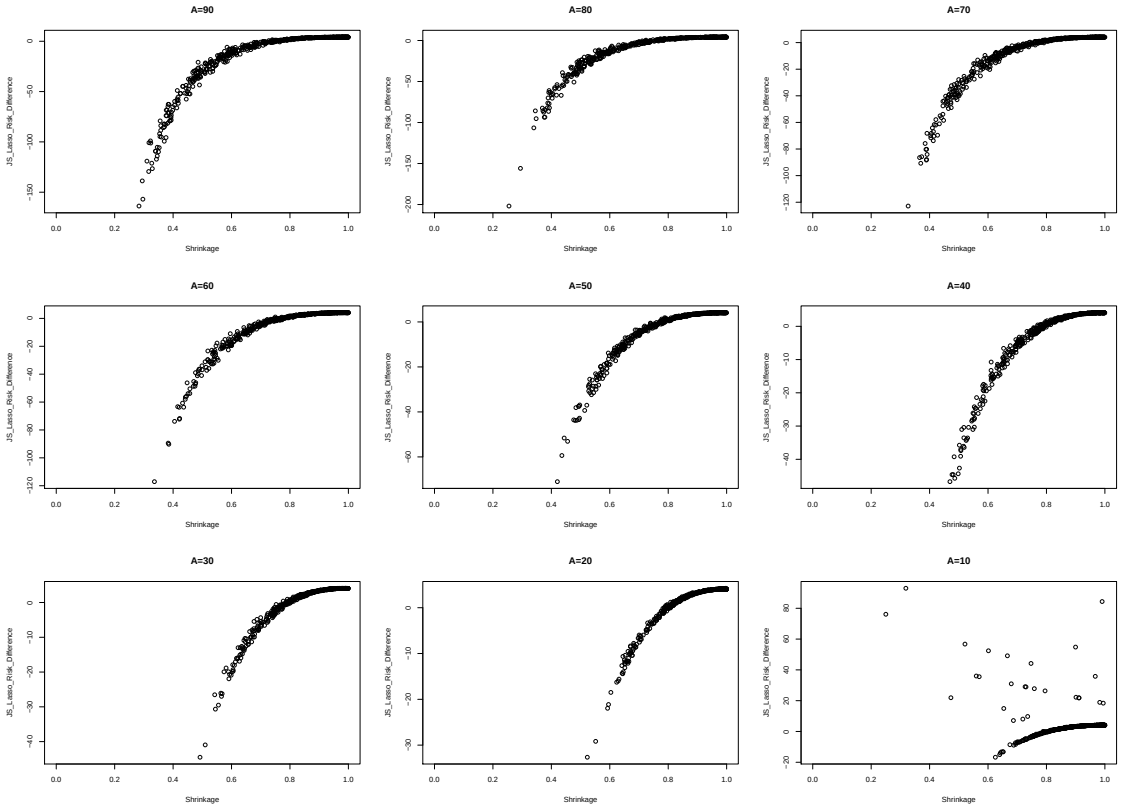

Figure 1.2 – $n = 200, p = 100, \sigma^2 = 9000$. Shrinkage Factor vs JS Estimate Risk Difference

Table 1.4 – JS estimate versus the Lasso for $n = 2000, p = 100, {\sigma^2}_\epsilon = 9000$

A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage	LossJS	LossLasso	LossOLS
90	-281.856	9415.62	9413.45	9444.88	0.4055	95.15	98.18	100.15
80	-48.712	9399.74	9394.85	9444.57	0.5446	92.53	93.62	99.04
70	-19.282	9395.25	9389.38	9460.70	0.6294	91.15	91.89	101.30
60	-9.530	9340.93	9335.16	9430.80	0.6758	84.50	84.97	100.07
50	-7.023	9316.86	9310.38	9435.23	0.7220	78.62	78.94	100.14
40	-2.978	9297.65	9289.44	9445.10	0.7756	70.47	70.86	100.14
30	-1.668	9271.85	9263.04	9467.94	0.8123	62.07	62.04	101.27
20	0.1828	9187.75	9175.19	9442.63	0.8509	50.37	50.24	99.96
10	1.634	9096.45	9078.83	9422.05	0.8866	34.57	33.62	99.61

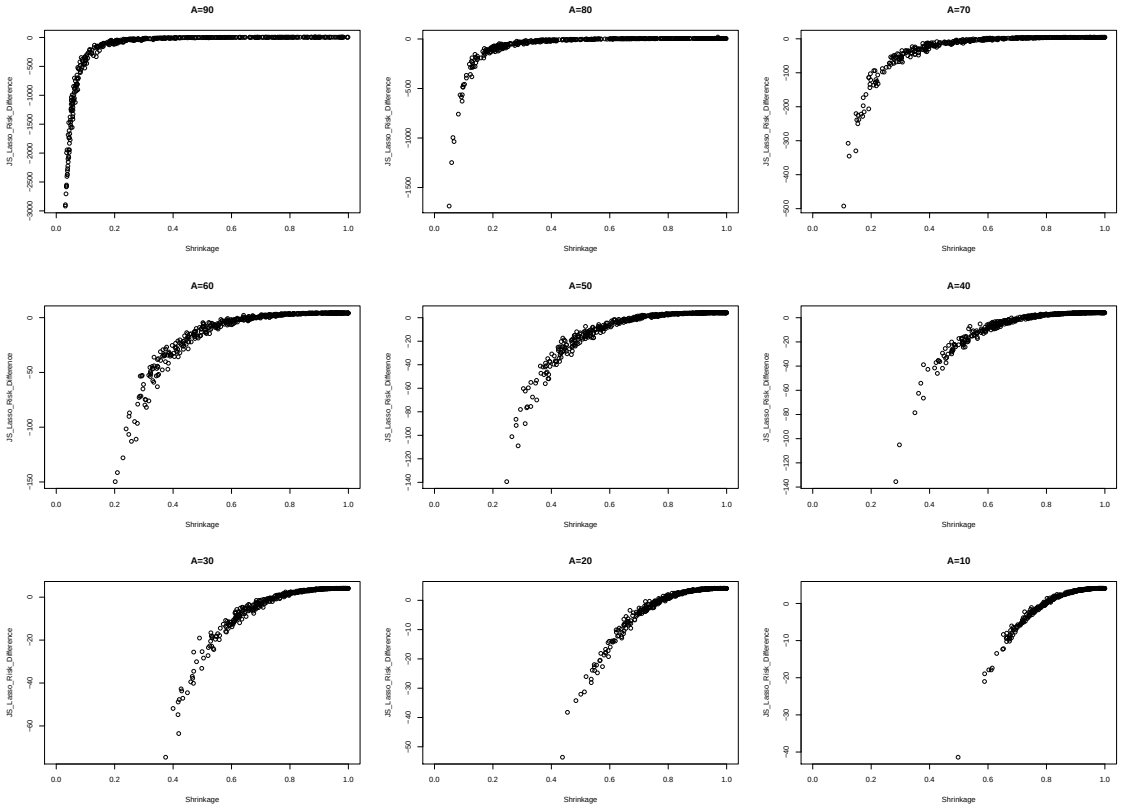

Figure 1.3 – $n = 2000, p = 100, \sigma^2 = 9000$. Shrinkage Factor vs JS Estimate Risk Difference

Table 1.5 – JS Estimate and Lasso for the diabetes data

p	A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage
10	7	-10.9152	2987.636	2978.308	2992.391	0.1393
11	10	-27.8007	2958.704	2954.805	2964.222	0.1184
12	11	-38.0976	2944.279	2939.618	2950.075	0.1102
13	12	-37.2379	2949.996	2946.905	2957.466	0.1264
14	10	-35.6415	2960.139	2957.266	2969.602	0.1441
15	10	1.5898	2961.256	2956.874	2973.513	0.3533
16	10	2.7291	2966.647	2956.874	2986.409	0.4711
17	10	3.5256	2969.958	2956.839	2998.502	0.5808
18	17	-2.2285	2942.761	2948.661	2961.343	0.3504
19	12	-45.0147	2911.783	2936.618	2928.498	0.1763
20	17	-26.0537	2828.888	2875.763	2843.930	0.2041
21	14	-24.6274	2837.269	2890.579	2855.139	0.2238
22	14	-18.2106	2845.097	2887.726	2865.896	0.2599
23	14	-17.1613	2851.165	2887.572	2875.333	0.2806
24	15	-19.8026	2835.457	2891.579	2860.599	0.2755
25	24	-18.4304	2843.393	2870.900	2872.245	0.2964
26	25	-18.3917	2843.505	2865.662	2875.236	0.3079
27	26	-18.4801	2853.542	2875.364	2885.187	0.3073

Table 1.6 – JS Estimate and Lasso for the diabetes data

p	A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage
28	27	-21.0438	2864.318	2885.237	2897.084	0.3009
29	28	-21.0009	2864.482	2891.879	2900.115	0.3110
30	29	-20.6710	2866.318	2895.675	2905.118	0.3226
31	30	-19.4684	2873.225	2909.543	2916.110	0.3396
32	31	-18.9295	2876.259	2919.591	2922.740	0.3521
33	24	-18.5834	2878.165	2925.410	2928.126	0.3629
34	32	-17.3397	2885.282	2925.873	2939.899	0.3801
35	32	-16.8797	2887.832	2924.589	2946.290	0.3914
36	25	-15.9764	2892.954	2936.251	2955.979	0.4060
37	25	-5.8806	2878.172	2916.149	2942.216	0.4866
38	25	-5.1835	2881.883	2915.912	2951.471	0.5083
39	25	-4.4633	2885.737	2915.724	2961.196	0.5286
40	25	-4.2182	2886.741	2915.144	2966.905	0.5404
41	25	-3.4474	2890.908	2915.052	2977.569	0.5619
42	25	-2.9410	2893.491	2914.748	2986.045	0.5786
43	25	-2.1513	2897.788	2914.743	2997.597	0.6017
44	15	-1.8604	2899.081	2917.295	3004.377	0.6140
45	15	-1.2044	2902.585	2917.253	3015.226	0.6351
46	15	-0.9564	2905.074	2918.422	3023.377	0.6465

Table 1.7 – JS Estimate and Lasso for the diabetes data

p	A	RiskDiff	PredJS	PredLasso	PredOLS	Shrinkage
47	15	-0.5756	2906.926	2918.236	3031.676	0.6612
48	15	-0.2484	2908.460	2918.026	3039.543	0.6747
49	28	-1.9131	2897.977	2914.081	3024.381	0.6381
50	28	-1.4902	2900.083	2913.774	3032.902	0.6521
51	15	-0.8434	2903.541	2916.345	3044.273	0.6717
52	15	-1.8325	2897.184	2915.281	3036.731	0.6528
53	15	-1.3294	2899.773	2915.184	3046.332	0.6680
54	15	-1.0088	2901.275	2914.980	3054.014	0.6792
55	15	-1.0012	2900.905	2914.572	3057.996	0.6830
56	15	-0.4228	2903.959	2914.562	3069.124	0.7006
57	15	0.1195	2906.810	2914.552	3080.240	0.7182
58	15	0.6276	2909.466	2914.542	3091.363	0.7357
59	15	0.3715	2907.507	2913.944	3091.727	0.7311
60	15	0.8317	2909.880	2913.915	3102.460	0.7475
61	15	1.2808	2912.200	2913.904	3113.509	0.7644
62	15	1.4686	2912.939	2913.683	3120.687	0.7732
63	15	1.4301	2912.300	2913.264	3124.144	0.7742
64	15	1.7933	2914.117	2913.219	3134.534	0.7893

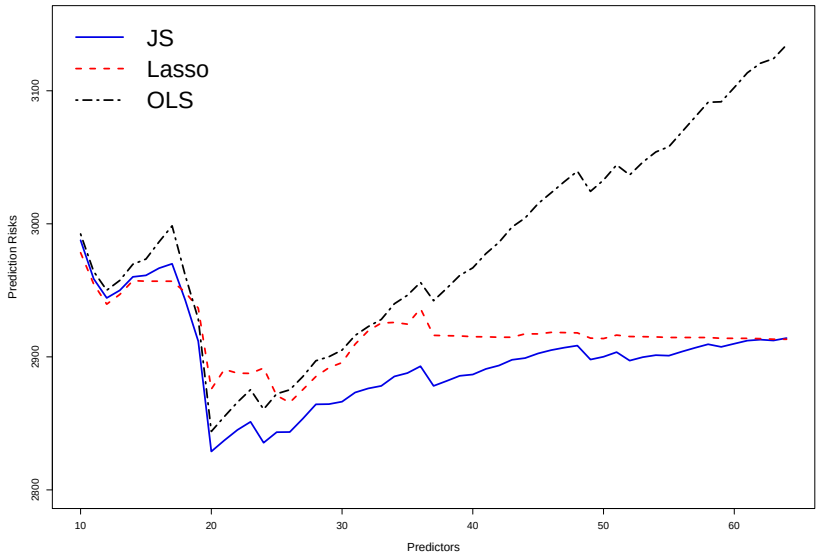

Figure 1.4 – Prediction Risk Comparisons for the Diabetes Data

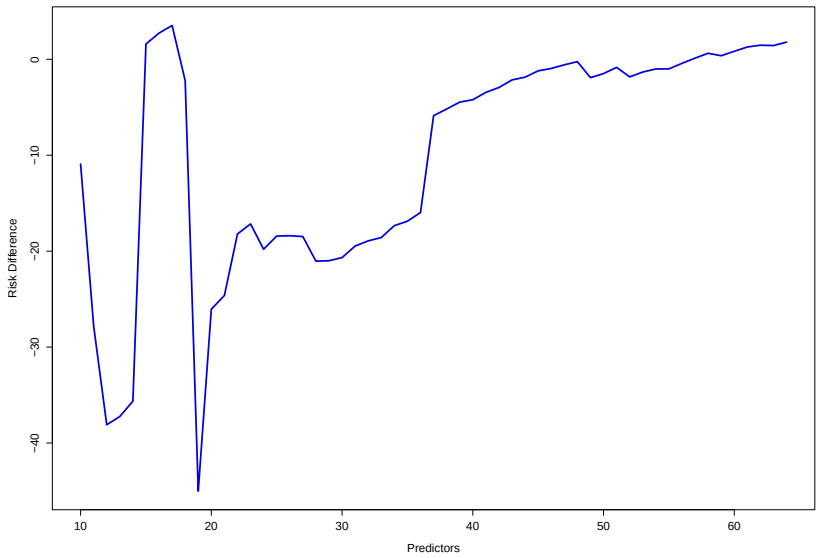

Figure 1.5 – Risk Difference between the Shrinkage Model and the Lasso

1.9 Estimators under a general constraint set

The constraint set under consideration here is the same crosspolytope as earlier. However, we consider the approach of treating the problem as considered by Kuriki and Takemura [30]. They treat the problem in great detail for the case when $X \sim N(\beta, I)$. We extend some of their results for an arbitrary covariance matrix Σ and observe that their results hold true in this more general setting.

Let us denote the constraint region by K that is a closed convex set in $\mathbb{R}^p$. We get the following theorem from [53].

Theorem 1.4. *Given a non-empty closed convex set K and a point $\mathbf{x}$ in $\mathbb{R}^p$, there exists a unique point $\mathbf{a_0}$ of K such that*

$$\|\mathbf{x} - \mathbf{a_0}\| = \inf\{\|\mathbf{x} - \mathbf{z}\| : \mathbf{z} \in \mathbf{K}\}$$

Thus, we can define a mapping $f : \mathbb{R}^p \to K$ as $f(\mathbf{x}) = \mathbf{a_0}$, where $\mathbf{a_0}$ is the closest point of K from $\mathbf{x}$. This mapping is the projection operator and the point $\mathbf{a_0}$ is the metric projection of $\mathbf{x}$ onto K_t. Further, this is a Lipschitz continuous map. For further details, refer [53].
Appealing to the Projection theorem, we have the following expression. For any vector $\mathbf{x}$ in $\mathbb{R}^p$ we get

$$\mathbf{x} = \mathbf{x_k} + (\mathbf{x} - \mathbf{x_k}) \tag{1.30}$$

Let ∂K denote the boundary of the set K. For a fixed $s \in \partial K$, the normal cone of K at s is given by $N(K, s) = \{y - s | y_K = s\}$.
We partition the boundary ∂K, depending on the dimension of the normal cone as

$$\partial K = D_1(\partial K) \cup \ldots \cup D_p(\partial K) \tag{1.31}$$

where

$$D_m(\partial K) = \{s \in \partial K | \dim N(K, s) = m\}$$

Define,

$$E_m(\partial K) = \{x \in \mathbb{R}^p \backslash K | x_K \in D_m(\partial K)\}$$

We make the following regularity condition on the smoothness of the boundary. $D_m(\partial K)$ is a $(p-m)$ dimensional C^2 manifold consisting of a finite number of relatively open connected components. If ∂K meets the above condition, it is called piecewise smooth.

We are dealing with the least squares estimates. $\hat{\beta}_0$ denotes the OLS estimate vector. Suppose that $\hat{\beta}_0 \notin K$ and let $s = \hat{\beta}_t \in D_m(\partial K)$ denote the projection of the OLS vector on the constraint set. Since $D_m(\partial K)$ is a $(p-m)$ dimensioanl C^2 manifold, there exists a local co-ordinate system given by $s = s(\theta), \theta = (\theta^1, \ldots, \theta^{p-m})$ in a neihbourhood of s. The tangent space $T_{s(\theta)}$ of $D_m(\partial K)$ at $s(\theta)$ is spanned by

$$\{b_a(\theta) = \frac{\partial s}{\partial \theta^a}, a = 1, \ldots, p-m\}$$

Tangent space $T_{s(\theta)}$ is a vector space and the above set forms a basis for the space. We can write the orthonormal basis of $T^{\perp}_{s(\theta)}$ as

$$\{n_\alpha(\theta), \alpha = 1, \ldots, m\}$$

and

$$\langle b_a(\theta), n_\alpha(\theta) \rangle = 0$$

The metric $G = G(\theta)$ of $D_m(\partial K)$ at $s = s(\theta)$ is

$$G(\theta) = (g_{ab}(\theta))_{1 \leq a,b \leq p-m}$$

and

$$g_{ab}(\theta) = \langle b_a(\theta), b_b(\theta) \rangle$$

Note: $T^{\perp}_{s(\theta)}$ is the affine hull of $N(K, s)$, we can write an element in $N(K, s)$ as $\sum_{\alpha=1}^{m} t^{\alpha} n_{\alpha}(\theta)$ with the new parameter $t = (t^1, \ldots, t^m)$. Hence the OLS vector can be written as

$$\hat{\beta}_0 = s(\theta) + n(\theta, t) \tag{1.32}$$

where $n(\theta, t) = \sum_{\alpha=1}^{m} t^{\alpha} n_{\alpha}(\theta)$. The Jacobian of this local one-to-one transformation, $x \leftrightarrow (\theta, t)$ was first given in the derivation of Weyl's tube formula as

Lemma 1.7.

$$dx = |I_{p-m} + H(\theta, t)| ds(\theta) dt \tag{1.33}$$

where $H(\theta, t)$ denotes the second fundamental form and

$$ds(\theta) = \sqrt{(G(\theta)} d\theta^1 \ldots d\theta^{p-m}$$

is the volume element of $D_m(\partial K)$, $dx = dx_1 \ldots dx_p$ and $dt = dt_1 \ldots dt_m$

We are projecting a vector in $\mathbb{R}^p$ onto a bounded polyhedron. Hence the second fundamental, $H(\theta, t) \equiv 0$. We now extend the results of Kuriki and Takemura [30] to the case of a non-orthogonal design. Using the above Jacobian, we can look at the distribution of (θ, t) from $\hat{\beta}_0$. In the rest of this section, we shall denote $\sigma^2 (X'X)^{-1}$ as Σ.

Lemma 1.8. *Let $\hat{\beta}_0 \sim N(\beta, \Sigma)$. The conditional density of $t = (t^1, \ldots, t^m)$ given $\beta_K = s(\theta) \in D_m(\partial K)$ is*

$$f(t|\theta) = e(\theta) \exp\{-\frac{1}{2} \|n(\theta, t)\|^2 + \langle n(\theta, t), \mu - s(\theta) \rangle\} dt$$

for t such that $n(\theta, t) \in N(K, s(\theta))$

Proof. We note from (1.32) that the random variable $\hat{\beta}_0$ is expressed as a sum of two components. Fixing the point of projection, $s(\theta)$ leaves $n(\theta, t)$ as the random component in that expression. Now,

$$f(\hat{\beta}_0)d\hat{\beta}_0 = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}} \exp[-\frac{1}{2}(\hat{\beta}_0 - \beta)'\Sigma^{-1}(\hat{\beta}_0 - \beta)]d\hat{\beta}_0.$$

Using the Jacobian from the Weyl tube formula and noting that $s(\theta)$ is assumed fixed for this calculation

$$\begin{aligned} f(t|\theta)dt &= C\exp[-\frac{1}{2}\|s(\theta) + n(\theta, t) - \beta\|^2]dt \\ &= C\exp[-\frac{1}{2}\|n(\theta, t) + (\beta - s(\theta)\|^2]dt \\ &= C\exp[-\frac{1}{2}\|n(\theta, t)\|^2 - \frac{1}{2}\|\beta - s(\theta)\|^2 + \langle n(\theta, t), \beta - s(\theta)\rangle]dt \\ &= e(\theta)\exp[-\frac{1}{2}\|n(\theta, t)\|^2 + \langle n(\theta, t), \beta - s(\theta)\rangle]dt. \end{aligned}$$

In the above expressions, C denotes the normalising constant independent of θ and $e(\theta)$ denotes the normalising constant as a function of θ. □

This result could be used to derive the distribution of the squared length of projection under smoothness conditions on the boundary of the convex set. Kuriki and Takemura[30] derive the conditional distribution of the squared length of projection, conditioned on the fixed point of projection, for the i.i.d. normal case. We conjecture a similar result for the arbitrary covariance case in the following lemma.

Lemma 1.9. *Let $\hat{\beta}_0 \sim N(\beta, \Sigma)$. Let the squared length of projection be given by*

$$l^2 = \|\hat{\beta}_0 - \hat{\beta}_t\|_\Sigma^2 \tag{1.34}$$

Substituting $u = l^{-1}t \in S^{m-1}$ we get the conditional density of l given $\hat{\beta}_t = s(\theta) \in D_m(\partial K)$

and u *such that* $n(\theta, u) \in N(K, s(\theta))$ *is*

$$f(l|\theta, u) = \begin{cases} e(\theta, u) \exp\{-\frac{1}{2}l^2 + l\langle n(\theta, u), \mu - s(\theta)\rangle\} l^{m-1} dl & \text{if } l \geq 0 \\ 0 & o.w. \end{cases} \tag{1.35}$$

Here, $e(\theta, u)$ *is a normalising constant as a function of* θ *and* u*.*

In Chapter 2, we empirically observe the conditional distribution of the squared length of projection, conditioned on the extent of sparsity in the model, to have a Weibull distribution.

Chapter 2

Loss Estimation

2.1 Introduction

Let X be a p-dimensional random vector having the distribution P_θ where $\theta \in \Theta \subset \mathbb{R}^p$ is a vector of unknown parameters. Given $d(X)$, an estimator of θ, it is common to analyse the quality of the estimation rule through a loss function $L(d(X), \theta)$. The loss function measures the discrepancy between the unknown parameter and its estimator. Typically, this loss is not known since θ is unknown. We derive conditions under which unbiased estimators of loss are dominated for the estimation of mean vector of a multivariate normal model with a known but arbitrary covariance matrix. Improved loss estimators are derived for the cases when the estimator of the unknown mean is chosen to be the MLE and an improved estimator, respectively. We further derive loss estimators that dominate the unbiased estimator of loss for θ for a linear model $Y = Z\theta + \epsilon$ when the estimation rule is the Lasso [45].

A global evaluation of the estimation procedure, based on all possible observations of the random X is usually obtained by looking at the risk, $R(d(X), \theta) = E_\theta[L(d(X), \theta)]$, where E_θ denotes expectation with respect to the distribution P_θ. Such a risk function can then be used to compare competing estimators of θ. Among various risk metrics, the maximum risk, $\bar{R}(d(X)) = \sup_\theta R(d(X), \theta)$ is often used as a frequentist assessment of risk. However, such a criterion is not data dependent because it averages over the sampling distribution of X and does not give due weight to the actual observation(s). As such, estimating the loss function directly using some function of the observed data would serve as a parameter estimation criterion that is conditional with respect to the observation(s) at hand. We thus

consider the problem of estimating the loss, using a data dependent estimator $\lambda(X)$. Such an estimator is known as the loss estimator. Thus, instead of using a global selection criterion, like the maximum risk or average risk, we choose a conditional approach and estimate the loss incorporating the information of the realisation of the random vector X.

To evaluate the performance of $\lambda(X)$ as an estimator of loss, we consider squared error

$$L^*(\lambda(X), \theta) = (\lambda(X) - L(d(X), \theta))^2.$$

In order to investigate the global performance criterion, we consider the expectation of the new loss function and call it the "mean distance" between $\lambda(X)$ and $L(d(X), \theta)$. It is expressed as the risk of $\lambda(X)$ as follows

$$\begin{aligned} \mathrm{M}(\lambda(X), L(d(X), \theta)) &= E_\theta[\lambda(X) - L(d(X), \theta)]^2 \\ &= E_\theta[\lambda(X) - \|d(X) - \theta\|^2]^2. \end{aligned}$$

Let $R(d, \theta) = E_\theta[L(d(X), \theta)]$ denote the risk of $d(X)$. Suppose $\lambda^U(X)$ is an unbiased estimator of risk, such that $E_\theta[\lambda^U(X)] = R(d, \theta)$. This implies $E_\theta[\lambda^U(X)] = E_\theta L(d(X), \theta)$ and thus $E_\theta[\lambda^U(X) - L(d(X), \theta)] = 0$. Consequently, we say that $\lambda^U(X)$ is an unbiased estimator of loss. Our objective is to exhibit conditions under which an estimator $\lambda(X)$ is such that it has smaller mean distance compared to an unbiased estimator of loss. In other words, we construct new estimators $\lambda(X)$ of $L(d(x), \theta)$ such that $\mathrm{M}(\lambda(X), L) \leq \mathrm{M}(\lambda^U(X), L)$.

An early work in this area was done by Sandved [38], where he proposed the notion of *best* unbiased estimators of loss. Johnstone [23] considered the problem of providing improved estimators of loss in an i.i.d. normal setting when the mean vector is estimated using both the MLE and the James-Stein estimator. He showed that if $X \sim N_p(\theta, I)$ and if we use $d(X) = X$ as the MLE of θ, then $\lambda^U(X) = p$ is an unbiased estimator of $L(X, \theta)$. He proved

that $\lambda(X) = p - 2(p-4)/\|X\|^2$ dominates $\lambda^U(X)$ when $p \geq 5$. He further showed that $\lambda(X) = p - (p-2)^2/\|X\|^2 + 2p/\|X\|^2$ improves upon the unbiased estimator of loss for the James-Stein estimator of θ. Wan and Zou [52] provided unbiased and improved estimators of loss for the unknown variance case in the i.i.d. normal setting. Rukhin [37] provided a loss function framework that combined the decision problem error with an inaccuracy estimate. In other words, such an estimator of loss gave a method for simultaneous reporting of decision and precision. In addition, he also provided necessary and sufficient conditions for admissibility of such a loss estimator. Lele [31] presents a unified method to prove admissibility of a variety of Bayes loss estimators for the general exponential family of distributions. Fourdrinier and Wells [21] estimate the loss of a point estimator for the case of spherically symmetric distributions for the general linear model. They present improved loss estimators for two different location estimators; the least squares estimator and a shrinkage estimator. More recently, Fourdrinier and Lepelletier [18] consider the problem of estimating $c(\|x-\theta\|^2)$ for a general nonnegative function c under usual quadratic loss. They provide a sufficient condition of domination over the unbiased estimator of loss in terms of a partial differential inequality.

An important application of the theory of loss estimation is for model selection. It is shown in Fourdrinier and Wells [20] that improved loss estimators give more accurate model selection procedures. Bartlett et al. [4] study model selection based on penalised empirical loss minimisation and indicate the relationship between loss estimation and data-based complexity penalisation. Any good loss estimate may be converted to a penalty function and the performance of the estimate is dictated by the quality of the estimate of loss. This establishes the relationship between complexity regularisation and good estimates of loss.

In Section 2.2, we consider estimating the loss for estimating the mean vector of X where $X \sim N_p(\theta, \Sigma)$, a p-dimensional multivariate normal vector with a known but arbitrary covariance matrix. Using a similar approach as Johnstone [23], we exhibit estimators of loss that dominate the unbiased estimator of loss. We provide a simulation study showing risk gains when the covariance matrix is a diagonal and a full ranked symmetric, positive definite matrix, respectively. In Section 2.3, we consider the same multivariate normal set-up, but now we seek to provide improved estimators of loss when an improved estimator is used to estimate θ. We give sufficient conditions for domination over an unbiased estimator of loss. In Section 2.4, we work in the linear model framework with a known error variance. We first construct an unbiased estimator of quadratic loss when Lasso is used to estimate the regression coefficients and we then exhibit an improved loss estimator that dominates Lasso under some conditions. Finally we present a detailed simulation study in Section 2.5 where risk gains of the improved loss estimator on using Lasso are shown for varying levels of sparsity.

2.2 Loss Estimation for MLE

Throught the chapter, we assume that the covariance matrix is completely specified. We state the following lemma from Fourdrinier, Strawderman and Wells [19] for the sake of completeness.

Lemma 2.1. *Let $X \sim N_p(\theta, \Sigma)$, where Σ is a positive definite symmetric matrix. Let $L(X, \theta) = \|X - \theta\|^2_\Sigma = (X - \theta)^T \Sigma^{-1} (X - \theta)$ denote the general quadratic loss function and $\langle \cdot, \cdot \rangle$ denote the inner product induced by this norm. Then, for a weakly differentiable function $g : \mathbb{R}^p \to \mathbb{R}^p$, we have, $E_\theta[\langle X - \theta, g(X) \rangle] = E_\theta[\text{div} g(X)]$, where $\text{div} g(X)$ denotes*

the divergence of $g(X)$.

Lemma 2.2. *Given* $X \sim N_p(\theta, \Sigma)$ *and a weakly differentiable function* $\gamma : \mathbb{R}^p \to \mathbb{R}$*, we have*

$$E_\theta[\|X-\theta\|_\Sigma^2 \gamma(X)] = E_\theta[p\,\gamma(X) + \text{div}\{\Sigma \nabla \gamma(X)\}].$$

Proof. Upon expanding the squared norm we have,

$$\|X-\theta\|_\Sigma^2 \gamma(X) = (X-\theta)^T \Sigma^{-1}(X-\theta)\gamma(X).$$

Letting $F(X) = \Sigma^{-1}(X-\theta)$ and $K(X) = F(X)\gamma(X)$, we get, $\|X-\theta\|_\Sigma^2 \gamma(X) = (X-\theta)^T K(X)$. Applying Lemma 2.1 we find,

$$\begin{aligned} E_\theta[\|X-\theta\|_\Sigma^2 \gamma(X)] &= E_\theta[(X-\theta)^T K(X)] \\ &= E_\theta[\text{div}(\Sigma K(X))] \\ &= E_\theta[\text{div}(\Sigma\Sigma^{-1}(X-\theta)\gamma(X))] \\ &= E_\theta[\text{div}\{(X-\theta)\gamma(X)\}]. \end{aligned}$$

We can now use the product rule of divergence as follows. Suppose $a = (a_1, a_2, \ldots, a_p)$ and $b = (b_1, b_2, \ldots, b_p)$ are two p-dimensional vectors and $a \cdot b = a_1 b_1 + a_2 b_2 + \ldots + a_p b_p$ denotes the dot product of two vectors a and b, then, given a vector F and a scalar valued function ϕ, we have

$$\text{div}(\phi F) = \nabla\phi \cdot F + \phi \text{div}(F). \tag{2.1}$$

Thus, $\text{div}(\gamma(X)(X-\theta)) = \nabla\gamma(X) \cdot (X-\theta) + \gamma(X)\text{div}(X-\theta)$. Since $\text{div}(X-\theta) = p$ and $\nabla\gamma(X) \cdot (X-\theta) = (X-\theta)^T \nabla\gamma(X)$, we have,

$$\begin{aligned} E_\theta[\|X-\theta\|_\Sigma^2 \gamma(X)] &= E_\theta[\text{div}\{(X-\theta)\gamma(X)\}] \\ &= E_\theta[(X-\theta)^T \nabla\gamma(X) + \gamma(X)p] \\ &= E_\theta[\text{div}\{\Sigma\nabla\gamma(X)\} + p\,\gamma(X)], \end{aligned}$$

where the last step, follows from Lemma 2.1. We thus have that an unbiased estimator of $\|X-\theta\|^2_\Sigma\gamma(X)$ is $p\,\gamma(X)+\text{div}\{\Sigma\nabla\gamma(X)\}$. □

Given a fixed estimator $d(X)$ of θ, define the invariant loss as $\|d(X)-\theta\|^2_\Sigma = (d(X)-\theta)^T\Sigma^{-1}(d(X)-\theta)$ where $X \sim N(\theta,\Sigma)$. If we use the MLE to estimate θ, then we choose $d(X)=X$. Denote the "improved" loss estimator by $\lambda(X)$ and let $\lambda^U(X)$ denote an unbiased estimator of $L(d(X),\theta)$. Then, the frequentist risk incurred by $\lambda(X)$ is $E_\theta[\lambda(X)-\|d(X)-\theta\|^2_\Sigma]^2$. It can be easily seen that when $d(X)=X$, the frequentist risk of $d(X)$ equals p. Therefore, $\lambda^U(X)=p$ is an unbiased estimator of $L(X,\theta)$ since $E[p-L(X,\theta)]=0$. An unbiased estimator for the loss is established if the estimation rule for θ is the MLE. The natural question would then be under what conditions can we provide improved estimators for loss.

Fixing the estimation rule to estimate θ to be the MLE, let us see how we can improve in estimating the loss. Consider improved loss estimators of the following form, $\lambda(X)=p-\gamma(X)$ where $\gamma:\mathbb{R}^p\to\mathbb{R}$. We need expressions for the expected distance or the risk incurred by $\lambda(X)$. The following theorem gives closed form expressions for $E_\theta[\lambda(X)-L(X,\theta)]^2$.

Theorem 2.1. *Suppose $X \sim N(\theta,\Sigma)$. Let $L(X,\theta)=(X-\theta)^T\Sigma^{-1}(X-\theta)$ denote the general quadratic loss function when X is used as an estimator of θ. Let $\lambda^U(X)$ be an unbiased estimator of loss for this problem and further suppose γ is a real valued function having two integrable weak derivatives. Then an unbiased estimator of the risk of $\lambda(X)$ is*

$$E_\theta[\lambda(X)-L(X,\theta)]^2 = E_\theta[2p+\gamma^2(X)+2\text{div}\{\Sigma\nabla\gamma(X)\}].$$

Proof. We know that $\lambda^U(X)=p$ is an unbiased estimator of $\|X-\theta\|^2_\Sigma$. Since $\lambda(X)$ is a real

valued function and given that $\gamma : \mathbb{R}^p \to \mathbb{R}$, we have,

$$\lambda(X) - \|X - \theta\|_\Sigma^2 = p - \gamma(X) - \|X - \theta\|_\Sigma^2.$$

Now,

$$\begin{aligned}[\|X - \theta\|_\Sigma^2 - \lambda(X)]^2 = [\|X - \theta\|_\Sigma^2 - p]^2 + \gamma^2(X) \\ + 2\gamma(X)[\|X - \theta\|_\Sigma^2 - p].\end{aligned} \tag{2.2}$$

Consider the first term in (2.2). Let $W = \|X - \theta\|_\Sigma^2$. Given $X \sim N(\theta, \Sigma)$, we get $W \sim \chi_p^2$. Therefore, $E_\theta(W) = p$ and $\text{Var}(W) = 2p$. Hence, $E_\theta[\|X - \theta\|_\Sigma^2 - p]^2 = \text{Var}(W) = 2p$. Taking expectations on both sides of (2.2) we get,

$$\begin{aligned}E_\theta[\|X - \theta\|_\Sigma^2 - \lambda(X)]^2 = E_\theta[\|X - \theta\|_\Sigma^2 - p]^2 + E_\theta[\gamma^2(X)] \\ + 2E_\theta\gamma(X)\|X - \theta\|_\Sigma^2 - 2pE_\theta\gamma(X).\end{aligned} \tag{2.3}$$

Using Lemma 2.2, we can write the third term in (2.3) as

$$E_\theta[\|X - \theta\|_\Sigma^2\gamma(X)] = E_\theta[p\gamma(X) + \text{div}\{\Sigma\nabla\gamma(X)\}]. \tag{2.4}$$

Therefore the right hand side of (2.2) equals

$$E_\theta[2p + \gamma^2(X) + 2\text{div}\{\Sigma\nabla\gamma(X)\}].$$

Thus, for the improved estimator $\lambda(X)$, we get an unbiased estimator of the mean distance to the loss function as

$$\text{M}(\lambda(X)) = 2p + \gamma^2(X) + 2\text{div}[\Sigma\nabla\gamma(X)]. \qquad \square$$

The risk of $\lambda^U(X)$ equals $2p$. Let $D(\theta, d, \lambda) = R(\theta, d, \lambda) - R(\theta, d, \lambda^U)$ denote the difference of the mean distance of the "improved" estimator $\lambda(X)$ and the unbiased estimator of loss $\lambda^U(X)$. Then choosing $\gamma(X)$ such that $\gamma^2(X) + 2\text{div}[\Sigma\nabla\gamma(X)] \leq 0$ with strict inequality on a set of positive measure gives a sufficient condition for the inadmissibility of $\lambda^U(X) = p$.

2.2.1 Examples

Let $X \sim N_p(\theta, \Sigma)$ where $\Sigma = D$ is a diagonal matrix with diagonal entries being $d_1, d_2, \ldots, d_p$. Let the squared norm of the observed vector x be $\|x\|^2_\Sigma = x^T\Sigma^{-1}x$. Choose $\gamma(x) = c/\|x\|^2_\Sigma$, where c is some constant. It can then be shown that

$$\gamma^2(x) + 2\text{div}(\Sigma\nabla\gamma(x)) = [c^2 - 4c(p-4)]\frac{1}{\|x\|^4_\Sigma}. \tag{2.5}$$

The right hand side of (2.5) is negative when $0 < c < 4(p-4)$. The greatest improvement happens for $c = 2(p-4)$ and this gives $\gamma^2(x) + 2\text{div}(\Sigma\nabla\gamma(x)) = -4(p-4)^2/\|x\|^4_\Sigma$. Consequently, we express the percentage risk gains of the new loss estimator as $100 * (R(\theta, d, \lambda^U) - R(\theta, d, \lambda))/R(\theta, d, \lambda^U)$ and this is tabulated in Table 2.1. For each simulation run, we generate a p-dimensional diagonal matrix Σ with entries $d_1, d_2, \ldots, d_p$ drawn independently from a uniform distribution with support on $(10, 40)$. Then we generate a p-dimensional multivariate normal random vector X as $X \sim N(0, \Sigma)$. The results in the second row of Table 2.1 present the percentage risk gains at $\theta = 0$ on using the improved estimator for loss over the unbiased estimator of loss when the estimation rule for θ is $d(X) = X$. Further, the results are indicated as the dimension of the random vector increases from $p = 10$ to $p = 100$ over the $10,000$ simulation runs for each such p.

As a second example, suppose $X \sim N(\theta, \Sigma)$ with Σ equal to a known but arbitrary covariance matrix. Generate a covariance matrix such that the correlation between X_i and X_j, the i^{th} and j^{th} components of X, is $\rho = 0.5^{|i-j|}$. We then generate a random vector X having a multivariate normal distribution with such a covariance structure. Define $\|x\|^2_\Sigma = x^T\Sigma^{-1}x$. Once again, choose $\gamma(x) = c/\|x\|^2_\Sigma$. Note that, for this choice of $\gamma(x)$, $\nabla\gamma(x) = -2c\Sigma^{-1}x/\|x\|^4_\Sigma$. It can be shown as before that the risk difference equals (2.5). The results in the third row of Table 2.1 show percentage risk gains at $\theta = 0$ on using the improved estimator for loss over

the unbiased estimator of loss for an arbitrary covariance matrix. The results are obtained over $10,000$ simulation runs as the dimension of the random vector increases from $p = 10$ to $p = 100$. It is interesting to note that the percentage of risk gains as shown in the Table 2.1 has the same order of magnitude as the results for the i.i.d. case presented in Johnstone [23]. We can thus deduce that the conditions of risk domination and the optimal choice of c for this choice of $\gamma(x)$ is invariant to the norm of x.

Table 2.1 –

Percent Risk Gains using improved loss estimator for diagonal and arbitrary covariance matrix

p	10	12	15	18	21	24	30	40	60	100
Diagonal Σ	14.75	13.06	11.34	9.69	8.53	7.60	6.22	4.74	3.19	1.95
Arbitrary Σ	14.79	13.27	11.29	9.66	8.50	7.50	6.14	4.70	3.20	1.97

2.3 Loss Estimation for Improved Estimators

The next natural step is the construction of improved estimators when the estimation rule used to estimate $\theta \in \Theta \subset \mathbb{R}^p$ is an improved estimator of the form $d(X) = X + g(X)$. Choosing $g(X) = (1 - (p-2)/\|X\|^2)X$ gives the classical James-Stein estimator. As before, suppose we observe X, a realisation from a p dimensional multivariate normal distribution with unknown vector of location parameter θ, and known positive definite covariance matrix Σ. Johnstone [23] further established the inadmissibility of unbiased estimators of loss if the estimation rule used to estimate θ is the improved estimator in the i.i.d. $N_p(\theta, I_p)$ setting.

Consider the estimator $d(X) = X + g(X)$ of θ, where $g : \mathbb{R}^p \to \mathbb{R}^p$ has two square integrable weak derivatives and consider the general quadratic loss function as $L(d(X), \theta) = (d(X) -$

$\theta)^T\Sigma^{-1}(d(X)-\theta)$. The risk function for this choice of $d(X)$ can be written as

$$\begin{aligned} E_\theta[L(d(X),\theta)] &= E_\theta[\|X+g(X)-\theta\|_\Sigma^2] \\ &= E_\theta[\|X-\theta\|_\Sigma^2 + \|g(X)\|_\Sigma^2 + 2\langle (X-\theta, g(X))\rangle] \\ &= p + E_\theta[\|g(X)\|_\Sigma^2 + 2\,\mathrm{div} g(X)]. \end{aligned} \tag{2.6}$$

The quantity inside the expectation operator on the right hand side of (2.6) is just a function of X, independent of θ. It is known as the Stein's Unbiased Risk Estimate (SURE), when $d(X)$ is of the form $X+g(X)$. As such, a natural unbiased estimator for the loss is SURE, which is,

$$\lambda^U(X) = p + \|g(X)\|_\Sigma^2 + 2\,\mathrm{div} g(X).$$

Stein [43] provided an unbiased estimator of the mean square distance between $\lambda^U(X)$ and $L(d(X),\theta)$ when $\Sigma = I_p$.

Theorem 2.2. *Suppose $X \sim N_p(\theta,\Sigma)$ and $\theta \in \Theta \subset \mathbb{R}^p$ and Σ is positive definite. Let $d(X) = X + g(X)$ be an estimator of θ and consider the loss function to be $L(d(X),\theta) = (d(X)-\theta)^T\Sigma^{-1}(d(X)-\theta)$. Let $\lambda^U(X) = p + \|g(X)\|_\Sigma^2 + 2\,\mathrm{div} g(X)$ be an unbiased estimator of the loss function $L(d(X),\theta)$ and $\lambda^g(X) = \lambda^U(X) - \gamma(X)$ be an improved estimator of the loss function, where $\gamma(X)$ is a twice differentiable real valued function. Then an unbiased estimator of the mean square distance between $\lambda^g(X)$ and $L(d(X),\theta)$ is given by*

$$\begin{aligned} M(\lambda^g,\theta) = 2p &+ 4E_\theta[\|h(X)\|^2 + (\mathrm{div} h(X))^2 + \mathrm{tr}\{\nabla h^T(X)\}^2 \\ &+ 2\sum_i\sum_j h_i(X)h_{jij}(X)] - 8E_\theta[\mathrm{div}\{g(X)\mathrm{div} g(X)\}] \\ &+ 4E_\theta[(\mathrm{div} g(X))^2] + 8E_\theta[\mathrm{div} g(X)] + E_\theta[\gamma^2(X)] \\ &+ 2E_\theta[\mathrm{div}\{\Sigma\nabla\gamma(X)\}] - 2E_\theta[\nabla\gamma(X)^T\mathrm{div} g(X)]. \end{aligned}$$

where $h(X) = \Sigma^{-1}g(X)$ and $h_{ij}(X) = \nabla_j h_i$ and $h_{iij} = \nabla_i\nabla_j h_i$.

Proof. We first need to extend Theorem 3 of Stein [43] for the case of a general symmetric positive definite Σ and further establish conditions for improved loss estimator when $d(X) = X + g(X)$.

Consider competing estimators that could serve to improve the unbiased loss estimator. Let $\lambda^g(X) = \lambda^U(X) - \gamma(X)$ be such a construction of an improved estimator of $L(X + g(X), \theta)$. The mean square distance is given by

$$\begin{aligned}
\mathrm{M}(\lambda^g, \theta) &= E_\theta[\lambda^g(X) - L(X + g(X), \theta)]^2 \\
&= E_\theta[\|X + g(X) - \theta\|_\Sigma^2 - \lambda^U(X) + \gamma(X)]^2 \\
&= E_\theta[\|X + g(X) - \theta\|_\Sigma^2 - p - \|g(X)\|_\Sigma^2 \\
&\qquad - 2\,\mathrm{div} g(X) + \gamma(X)\}]^2. \\
&= E_\theta[A + \gamma(X)]^2 \\
&= E_\theta[A^2 + 2A\gamma(X) + \gamma^2(X)]. \qquad (2.7)
\end{aligned}$$

where $A = \|X + g(X) - \theta\|_\Sigma^2 - p - \|g(X)\|_\Sigma^2 - 2\,\mathrm{div} g(X)$. Let us look at each term in (2.7) separately. The first term in the content of the expectation in (2.7) is

$$\begin{aligned}
A^2 &= [\|X + g(X) - \theta\|_\Sigma^2 - p - \|g(X)\|_\Sigma^2 - 2\,\mathrm{div} g(X)]^2 \\
&= [\|X - \theta\|_\Sigma^2 + 2\langle (X - \theta, g(X)\rangle - p - 2\,\mathrm{div} g(X)]^2 \\
&= [\|X - \theta\|_\Sigma^2 - p]^2 + 4[\langle (X - \theta), g(X)\rangle - \mathrm{div} g(X)]^2 \\
&\qquad + 4[\{\|X - \theta\|_\Sigma^2 - p\}\{\langle (X - \theta), g(X)\rangle - \mathrm{div} g(X)\}]. \qquad (2.8)
\end{aligned}$$

Once again, we need to look at (2.8) term by term. We know that $E_\theta[\|X - \theta\|_\Sigma^2 - p]^2 = 2p$. Thus, the expectation of the first term (2.8) is $2p$. Let's look at the second term in (2.8).

Denote the second term in (2.8) by A_2.

$$\begin{aligned} E_\theta[A_2] &= 4E_\theta[\langle (X-\theta), g(X)\rangle - \text{div}g(X)]^2 \\ &= 4E_\theta[\langle (X-\theta), g(X)\rangle^2 - 2\langle (X-\theta), g(X)\rangle \text{div}g(X) + \{\text{div}g(X)\}^2]. \end{aligned} \tag{2.9}$$

Consider the first term in (2.9), $E_\theta[\langle (X-\theta), g(X)\rangle^2] = E_\theta[(X-\theta)^T\Sigma^{-1}g(X)]^2$. Since Σ^{-1} is a matrix of constants, letting $h(X) = \Sigma^{-1}g(X)$, we get $E_\theta[\langle (X-\theta), g(X)\rangle^2] = E_\theta[(X-\theta)^T h(X)]^2$. This has the same form as Theorem 3 in Stein [43]. Thus, the expression becomes,

$$\begin{aligned} E_\theta[\langle (X-\theta), g(X)\rangle^2] = E_\theta[\|h(X)\|^2 + \{\text{div}h(X)\}^2 + \text{tr}\{\nabla h^T(X)\}^2 \\ + 2\sum_i \sum_j h_i(X) h_{jij}(X)]. \end{aligned} \tag{2.10}$$

Thus, we see that the expectation of the first term in (2.9) which was a function of X and θ can be written just as a function of X as in (2.10). Now, consider the second term inside the expectation on the right hand side of (2.9). Note that $\text{div}g(X)$ is just a scalar function of X, independent of θ. Thus, applying Lemma 2.1 to the second term within the expectation we get,

$$E_\theta[\langle (X-\theta), g(X)\rangle \text{div}g(X)] = E_\theta[\text{div}\{g(X)\text{div}(g(X))\}]. \tag{2.11}$$

The last term inside the expectation of (2.9) being just a function of X is fine. Hence, collecting the terms from (2.10) and (2.11) we get an expression for (2.9) just as a function of X, independent of θ. We next need to look at the third term in (2.8). Let's call this A_3.

Thus,

$$\begin{aligned}\frac{A_3}{4} &= \{\|X-\theta\|_\Sigma^2 - p\}\{\langle (X-\theta), g(X)\rangle - \mathrm{div} g(X)\} \\ &= [\{\|X-\theta\|_\Sigma^2\}\{\langle (X-\theta), g(X)\rangle\} - \{\|X-\theta\|_\Sigma^2\}\{\mathrm{div} g(X)\} \\ &\quad - p\,\langle (X-\theta), g(X)\rangle + p\,\mathrm{div} g(X)] \\ &= A_{31} - A_{32} - A_{33} + A_{34}.\end{aligned}$$

Now consider A_{31}. Taking expectation of this term we have,

$$\begin{aligned}E_\theta[A_{31}] &= E_\theta\{\|X-\theta\|_\Sigma^2\}\{\langle (X-\theta), g(X)\rangle\}. \\ &= E_\theta[\mathrm{div}(\alpha g(X))].\end{aligned}$$

where $\alpha = \|X-\theta\|_\Sigma^2$ and so the content of the above expectation term is not indepedent of θ. Applying the product rule of divergence as in (2.1), we have, $\mathrm{div}(\alpha g(X)) = (\nabla\alpha)^T g(X) + \alpha\,\mathrm{div} g(X)$. In our case, $\nabla\alpha = \nabla[\|X-\theta\|_\Sigma^2] = 2\Sigma^{-1}(X-\theta)$, so, $(\nabla\alpha)^T g(X) = 2(X-\theta)^T\Sigma^{-1} g(X)$. Therefore, $E_\theta[\mathrm{div}\{\alpha g(X)\}] = 2E_\theta[\langle (X-\theta), g(X)] + E_\theta[\alpha\,\mathrm{div} g(X)]$. Using Lemma 2.1 we thus get,

$$E_\theta[\mathrm{div}\{\alpha g(X)\}] = 2E_\theta[\mathrm{div} g(X)] + E_\theta[\alpha\,\mathrm{div} g(X)].$$

Therefore we have,

$$\begin{aligned}E_\theta[A_{31}] &= 2E_\theta[\mathrm{div} g(X)] + E_\theta[\alpha\, g(X)] \\ &= 2E_\theta[\mathrm{div} g(X)] + E_\theta[A_{32}].\end{aligned}$$

Consequently, $E_\theta[A_3] = 8E_\theta[\mathrm{div} g(X)] - 4E_\theta[A_{33}] + 4E_\theta[A_{34}]$. Note that $E_\theta[A_{32}]$ cancels out in the above expression since $E_\theta[A_{33}] = E_\theta[p\,\langle (X-\theta), g(X)\rangle] = E_\theta[p\,\mathrm{div} g(X)]$. Note also that $E[A_{33}]$ cancels with $E[A_{34}]$, which leaves

$$E_\theta[A_3] = 8E_\theta[\mathrm{div} g(X)]. \tag{2.12}$$

Combining all the above terms, we see that $E[A^2]$ can be written just as a function of X, independent of θ. Now, we need an expression for the second term in (2.7). The second term of (2.7) is

$$\begin{aligned} E_\theta[A\gamma(X)] &= E_\theta[\|X + g(X) - \theta\|_\Sigma^2 - p - \|g(X)\|_\Sigma^2 - 2\,\mathrm{div} g(X)]\gamma(X) \\ &= E_\theta[\|X - \theta\|_\Sigma^2 + 2\langle (X-\theta), g(X)\rangle - 2\,\mathrm{div} g(X)]\gamma(X) \\ &= E_\theta[\|X - \theta\|_\Sigma^2]\gamma(X) + 2E_\theta[\langle (X-\theta), g(X)\gamma(X)\rangle] \\ &\quad - 2E_\theta[\gamma(X)\mathrm{div} g(X)] - E_\theta[p\,\gamma(X)]. \end{aligned}$$

Applying Lemma 2.2, we get expressions for the first term in terms of X and using Lemma 2.1 we can get an expression for the second term purely as a function of X. The third term above is already a function of X so nothing needs to be done for that. Consequently,

$$E_\theta[A\gamma(X)] = E_\theta[\mathrm{div}\{\Sigma\nabla\gamma(X)\}] - 2E_\theta[\nabla\gamma(X)^T \mathrm{div} g(X)]. \tag{2.13}$$

Combining expressions for (2.9), (2.12) and (2.13), we can write (2.7) as

$$\mathrm{M}(\lambda^g, \theta) = E[A^2 + 2A\gamma(X) + \gamma^2(X)]. \qquad \square$$

All the terms in the right hand side of the above equation are just functions of X, independent of θ. As such, we can write the sufficient condition for improved estimator of loss corresponding to an improved estimator of θ as $\mathrm{M}(\lambda^g, \theta) \leq \mathrm{M}(\lambda^U(X), \theta)$.

2.4 Loss Estimation for the Lasso

Extending the ideas of risk domination from Sections 2.2 and 2.3, in this section, we will explore the domination in the context of the Lasso. In Theorem 2.3 in Subsection 2.4.1, we derive an unbiased estimator of the risk of the improved estimator of the loss difference between Lasso and least squares for the linear model. In Subsection 2.4.2, we derive sufficient

conditions under which we can get improved estimators for loss when Lasso is used for model selection. In Subsection 2.4.3, we present an example satisfying the sufficient conditions for domination of loss when Lasso is used.

2.4.1 Risk Difference Expression

Consider a linear model, $Y = Z\theta + \epsilon$, where $Y^{n\times 1}$ is a response vector, $Z^{n\times p}$ is a matrix having n observations on p covariates, $\theta^{p\times 1}$ is the unknown vector of regression coefficients and $\epsilon \sim N(0, \sigma^2 I_n)$. The Lasso estimate is obtained by solving

$$\underset{\theta}{\operatorname{argmin}} \|y - Z\theta\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\theta_j| \leq t. \tag{2.14}$$

The equivalent Lagrange multiplier formulation is given by

$$\underset{\theta}{\operatorname{argmin}} \|y - Z\theta\|_2^2 + \lambda \sum_{j=1}^{p} |\theta_j| \tag{2.15}$$

where t and λ are equivalent nonnegative tuning parameters. Let $K_t := \{\theta : \sum_{j=1}^{p} |\theta_j| \leq t\}$. We know from Kato [27] that Lasso is equivalent to the projection of the least squares estimator onto the closed convex set K_t under the general quadratic norm, i.e., if v_1 and v_2 are two p-dimensional vectors, then consider the norm induced by the inner product between the two vectors as $\langle v_1, v_2 \rangle = v_1^T \Sigma v_2$, where $\Sigma = (Z^T Z)/\sigma^2$.

Let $\hat{\theta}_0$ denote the ordinary least squares estimator and let $\hat{\theta}_K$ denote the Lasso estimator obtained by solving (2.14) or (2.15). Both $\hat{\theta}_0$ and $\hat{\theta}_K$ are estimators of the regression coefficient θ. For the following calculations, let us assume the tuning parameter to be fixed and so the constraint set K_t is denoted by K. It is well known that $\hat{\theta}_0 \sim N(\theta, \Sigma)$ where $\Sigma = \sigma^2 (Z^T Z)^{-1}$. Define a loss function as $L(d, \theta) = \|d - \theta\|_\Sigma^2$.

Define $\Delta L = L(d_2, \theta) - L(d_1, \theta)$ where $d_1 = \hat{\theta}_0$ and $d_2 = \hat{\theta}_K$. Thus, the difference in risks is given by $\Delta R = E_\theta[L(d_2, \theta)] - E_\theta[L(d_1, \theta)]$. Let λ^U denote an unbiased estimator of the

risk difference, that is, $E_\theta[\lambda^U] = \Delta R$. Let $l^2 = \|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2$ and m denote the codimension of the face of the projection of the least squares estimator onto the Lasso constraint set. In other words, m denotes the number of zeroes as estimated by Lasso for a given tuning parameter. From Theorem 1.1, we know that $\lambda^U(\hat{\theta}_0) = l^2 - 2m$ is an unbiased estimator of the difference in losses when $\hat{\theta}_0$ and $\hat{\theta}_K$ are used as estimators of θ respectively. Consider "improved" estimator of the loss difference as

$$\lambda^g(\hat{\theta}_0) = \lambda^U(\hat{\theta}_0) - \gamma(\hat{\theta}_0)$$

where $\gamma : \mathbb{R}^p \to \mathbb{R}$ has two integrable weak derivatives. Thus, the mean squared distance between $\lambda^g(\hat{\theta}_0)$ and ΔL is

$$E_\theta[\lambda^g(\hat{\theta}_0) - \Delta L]^2 = E_\theta[\lambda^U(\hat{\theta}_0) - \gamma(\hat{\theta}_0) - \Delta L]^2. \tag{2.16}$$

As a result, we have the following theorem.

Theorem 2.3. *Let ΔL denote the loss difference between Lasso and the ordinary least squares estimator for the coefficients of a linear model. Let $\lambda^g(\hat{\theta}_0)$ denote an improved estimator over $\lambda^U(\hat{\theta}_0)$, an unbiased estimator of ΔL. Then, $M(\lambda^g, \Delta L)$, the mean squared distance between $\lambda^g(\hat{\theta}_0)$ and ΔL is given by*

$$\begin{aligned} E_\theta[\lambda^g - \Delta L]^2 = 4E_\theta[\|h(\hat{\theta}_0)\|^2 + (\mathrm{div}h(\hat{\theta}_0))^2 + \mathrm{tr}\{\nabla h^T(\hat{\theta}_0)\}^2 \\ + 2\sum_i \sum_j h_i(\hat{\theta}_0) h_{jij}(\hat{\theta}_0) + m^2 - 2m\,\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K) \\ + m\gamma(\hat{\theta}_0) - \mathrm{div}\{(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)\}] + E_\theta[\gamma^2(\hat{\theta}_0)], \end{aligned}$$

where $\gamma(\hat{\theta}_0)$ is a real valued function and $h(\hat{\theta}_0) = \Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K)$. $h_{ij} = \nabla_j h_i$ and $h_{iij} = \nabla_i \nabla_j h_i$.

Proof. Rewriting the right hand side in (2.16) we have,

$$E_\theta[\lambda^g - \Delta L]^2 = E_\theta[\{\lambda^U(\hat{\theta}_0) - \Delta L\}^2 - 2\gamma(\hat{\theta}_0)\{\lambda^U(\hat{\theta}_0) - \Delta L\} + \gamma^2(\hat{\theta}_0)]. \tag{2.17}$$

Note that the second term in (2.17) equals $E_\theta[\gamma(\hat{\theta}_0)\lambda^U(\hat{\theta}_0)] - E_\theta[\gamma(\hat{\theta}_0)\Delta L]$. Hence,

$$\begin{aligned} E_\theta[\gamma(\hat{\theta}_0)\Delta L] &= E_\theta[\{\|\hat{\theta}_K - \theta\|_\Sigma^2 - \|\hat{\theta}^0 - \theta\|_\Sigma^2\}\gamma(\hat{\theta}_0)] \\ &= E_\theta[\{\|\hat{\theta}_K - \hat{\theta}_0 + \hat{\theta}_0 - \theta\|_\Sigma^2 \\ &\quad - \|\hat{\theta}^0 - \theta\|_\Sigma^2\}\gamma(\hat{\theta}_0)] \\ &= E_\theta[\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2\gamma(\hat{\theta}_0)] \\ &\quad - 2E_\theta[\langle\hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K\rangle\gamma(\hat{\theta}_0)]. \end{aligned} \tag{2.18}$$

Letting $g(\hat{\theta}_0) = (\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)$ we have,

$$\begin{aligned} E_\theta[\langle\hat{\theta}_0 - \theta, g(\hat{\theta}_0)\rangle] &= E_\theta[(\hat{\theta}_0 - \theta)^T\Sigma^{-1}g(\hat{\theta}_0)] \\ &= E_\theta[\mathrm{div}g(\hat{\theta}_0)] \\ &= E_\theta[\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)]. \end{aligned}$$

Therefore, it follows that,

$$\begin{aligned} E_\theta[\gamma(\hat{\theta}_0)(\lambda^U(\hat{\theta}_0) - \Delta L)] &= E_\theta[\gamma(\hat{\theta}_0)\lambda^U(\hat{\theta}_0)] - E[\gamma(\hat{\theta}_0)\Delta L] \\ &= E_\theta[\gamma(\hat{\theta}_0)\lambda^U(\hat{\theta}_0) - \|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2\gamma(\hat{\theta}_0) \\ &\quad + 2\mathrm{div}\{(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)\}]. \end{aligned}$$

Substituting $\lambda^U(\hat{\theta}_0) = \|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 - 2m$ we have,

$$\begin{aligned} &= E_\theta[\gamma(\hat{\theta}_0)\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 - 2m\gamma(\hat{\theta}_0) \\ &\quad - \gamma(\hat{\theta}_0)\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 + 2\mathrm{div}\{(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)\}] \\ &= E_\theta[-2m\gamma(\hat{\theta}_0) + 2\mathrm{div}\{(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)\}]. \end{aligned} \tag{2.19}$$

Consider the first term in (2.17),

$$\begin{aligned}
E_\theta[\lambda^U(\hat{\theta}_0) - \Delta L]^2 &= E_\theta[\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 - 2m - \Delta L]^2 \\
&= E_\theta[\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 - 2m - \|\hat{\theta}_K - \theta\|_\Sigma^2 + \|\hat{\theta}^0 - \theta\|_\Sigma^2]^2 \\
&= E_\theta[\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 - 2m - \|\hat{\theta}_K - \hat{\theta}_0 + \hat{\theta}_0 - \theta\|_\Sigma^2 \\
&\quad + \|\hat{\theta}^0 - \theta\|_\Sigma^2]^2 \\
&= E_\theta[\|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 - 2m - \|\hat{\theta}^0 - \theta\|_\Sigma^2 - \|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 \\
&\quad + 2\langle\hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K\rangle + \|\hat{\theta}^0 - \theta\|_\Sigma^2]^2 \\
&= E_\theta[2\langle\hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K\rangle - 2m]^2 \\
&= 4E_\theta[\langle\hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K\rangle^2 + m^2 - 2m\langle\hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K\rangle]. \qquad (2.20)
\end{aligned}$$

Again, letting $g(\hat{\theta}_0) = (\hat{\theta}_0 - \hat{\theta}_K)$, we can write the first term in (2.20) as,

$$E_\theta[\langle(\hat{\theta}_0 - \theta), g(\hat{\theta}_0)\rangle^2] = E_\theta[(\hat{\theta}_0 - \theta)^T\Sigma^{-1}g(\hat{\theta}_0)]^2.$$

Since Σ^{-1} is a matrix of constants, let $h(\hat{\theta}_0) = \Sigma^{-1}g(\hat{\theta}_0)$. Therefore,

$$E_\theta[\langle\hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K\rangle^2] = E_\theta[(\hat{\theta}_0 - \theta)^T h(\hat{\theta}_0)]^2.$$

This has the same form as Theorem 3 in Stein [43]. Hence,

$$\begin{aligned}
&= E_\theta[\|h(\hat{\theta}_0)\|^2 + (\mathrm{div}h(\hat{\theta}_0))^2 + \mathrm{tr}\{\nabla h^T(\hat{\theta}_0)\}^2 \\
&\quad + 2\sum_i\sum_j h_i(\hat{\theta}_0)h_{jij}(\hat{\theta}_0)]. \qquad (2.21)
\end{aligned}$$

As such, we can write $E[\lambda^U(\hat{\theta}_0) - \Delta L]^2$ as a function of $\hat{\theta}_0$, independent of θ, that is,

$$\begin{aligned}
E_\theta[\lambda^U(\hat{\theta}_0) - \Delta L]^2 &= 4E_\theta[[\|h(\hat{\theta}_0)\|^2 + \{\mathrm{div}h(\hat{\theta}_0)\}^2 + \mathrm{tr}\{\nabla h^T(\hat{\theta}_0)\}^2 \\
&\quad + 2\sum_i\sum_j h_i(\hat{\theta}_0)h_{jij}(\hat{\theta}_0) + m^2 - 2m\,\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K)].
\end{aligned}$$

Combining (2.17), (2.19) and (2.21), we get an expression for the squared distance between the improved loss estimator of the loss difference of Lasso and least squares estimator as follows

$$\begin{aligned} E_\theta[\lambda^g - \Delta L]^2 &= E[{}_\theta\lambda^U(\hat{\theta}_0) - \Delta L - \gamma(\hat{\theta}_0)]^2 \\ &= 4E_\theta[[\|h(\hat{\theta}_0)\|^2 + \{\mathrm{div}h(\hat{\theta}_0)\}^2 + \mathrm{tr}\{\nabla h^T(\hat{\theta}_0)\}^2 \\ &\quad + 2\sum_i \sum_j h_i(\hat{\theta}_0)h_{jij}(\hat{\theta}_0) + m^2 - 2m\,\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K) \\ &\quad + m\gamma(\hat{\theta}_0) - \mathrm{div}\{(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)\}] + E_\theta[\gamma^2(\hat{\theta}_0)]. \end{aligned} \tag{2.22}$$

□

We thus note that all the terms inside the expectation operator in the right hand side of (2.22) are independent of θ. Consequently, we get an expression for the mean square distance between the improved loss estimator and the difference in loss functions due to Lasso and least squares only as the function of the observations. Choosing $\gamma(\hat{\theta}_0)$ such that $\mathrm{M}(\lambda^g(\hat{\theta}_0), \Delta L) \leq \mathrm{M}(\lambda^U(\hat{\theta}_0), \Delta L)$ yields a sufficient condition for improved estimators of the loss difference between Lasso and ordinary least squares.

2.4.2 Conditions for Improved Loss Estimators for Lasso

Here, we establish the sufficient conditions of domination over an unbiased estimator of loss when the estimator of the regression coefficients in a linear model is the Lasso. As in Subsection 2.4.1, we consider $Y = Z\theta + \epsilon$ to represent the linear model in question where $\epsilon \sim N(0, \sigma^2 I_n)$. We assume σ^2 is known.

Define the loss function of doing Lasso by $L(\hat{\theta}_K, \theta) = \|\hat{\theta}_K - \theta\|_\Sigma^2$. Denote an unbiased estimator of $L(\hat{\theta}_K, \theta)$ by $\lambda^U(X)$. As a simple corollary to Theorem 1.1, we can say that $\lambda^U(\hat{\theta}_0) = l^2 - 2m + p$ where $l^2 = \|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2$ and m is the number of zeroes in the Lasso

solution. Consider estimators of the form $\lambda^g(\hat{\theta}_0) = \lambda^U(\hat{\theta}_0) - \gamma(\hat{\theta}_0)$, where $\gamma : \mathbb{R}^p \to \mathbb{R}$ has two weak derivatives. Let $R(\lambda^U(\hat{\theta}_0), L)$ and $R(\lambda^g(\hat{\theta}_0), L)$ denote the mean risk of the unbiased estimator of loss and the "improved" estimator of loss, respectively. Then it follows that,

$$\begin{aligned} D(\lambda^g, \lambda^U, L) &= R(\lambda^g(\hat{\theta}_0), L) - R(\lambda^U(\hat{\theta}_0), L) \\ &= E_\theta[\gamma^2(\hat{\theta}_0)] - 2E_\theta[(\lambda^U(\hat{\theta}_0) - L)\gamma(\hat{\theta}_0)] \\ &= E_\theta[\gamma^2(\hat{\theta}_0)] - 2E_\theta[(l^2 - 2m + p)\gamma(\hat{\theta}_0)] + 2E_\theta[\|\hat{\theta}_K - \theta\|_\Sigma^2 \gamma(\hat{\theta}_0)]. \end{aligned} \tag{2.23}$$

The first two terms in the right hand side of (2.23) are not dependent on θ but the third term is dependent on both the estimator and θ. Consider the third term in (2.23).

$$\begin{aligned} E_\theta[\|\hat{\theta}_K - \theta\|_\Sigma^2 \gamma(X)] &= E_\theta[\|\hat{\theta}_K - \hat{\theta}_0 + \hat{\theta}_0 - \theta\|_\Sigma^2 \gamma(X)] \\ &= E_\theta[\{\|\hat{\theta}^0 - \theta\|_\Sigma^2 + \|\hat{\theta}_0 - \hat{\theta}_K\|_\Sigma^2 \\ &\qquad - 2\langle \hat{\theta}_0 - \theta, \hat{\theta}_0 - \hat{\theta}_K \rangle\}\gamma(\hat{\theta}_0)] \\ &= E_\theta[l^2\gamma(\hat{\theta}_0) + \|\hat{\theta}^0 - \theta\|_\Sigma^2 \gamma(\hat{\theta}_0)] \\ &\qquad - 2E_\theta[\mathrm{div}\{(\hat{\theta}_0 - \hat{\theta}_K)\gamma(\hat{\theta}_0)\}]. \end{aligned} \tag{2.24}$$

From Lemma 2.2 we know,

$$E_\theta[\|\hat{\theta}^0 - \theta\|_\Sigma^2 \gamma(\hat{\theta}_0)] = E_\theta[p\gamma(\hat{\theta}_0) + \mathrm{div}\{\Sigma \nabla \gamma(\hat{\theta}_0)\}].$$

Furthermore it follows that,

$$\mathrm{div}\{\hat{\theta}_0 - \hat{\theta}_K\}\gamma(\hat{\theta}_0) = \nabla\gamma(\hat{\theta}_0)^T(\hat{\theta}_0 - \hat{\theta}_K) + \gamma(\hat{\theta}_0)\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K).$$

As an extension based on Zou et al. [58] we get, $\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K) = m$, hence,

$$\begin{aligned} E_\theta[\|\hat{\theta}_K - \theta\|_\Sigma^2 \gamma(X)] = E_\theta[(l^2 + p)\gamma(\hat{\theta}_0) + \mathrm{div}\{\Sigma\nabla\gamma(\hat{\theta}_0)\}] \\ - 2E_\theta[\nabla\gamma(\hat{\theta}_0)^T(\hat{\theta}_0 - \hat{\theta}_K) + m\gamma(\hat{\theta}_0)]. \end{aligned} \tag{2.25}$$

Therefore we can express (2.23) as

$$D(\lambda^g, \lambda^U, L) = E_\theta[\gamma^2(\hat{\theta}_0) + 2\text{div}(\Sigma\nabla\gamma(\hat{\theta}_0)) - 4\nabla\gamma(\hat{\theta}_0)^T(\hat{\theta}_0 - \hat{\theta}_K)]. \tag{2.26}$$

Thus a sufficient condition for getting improved loss estimators for Lasso is

$$\gamma^2(\hat{\theta}_0) + 2\,\text{div}(\Sigma\nabla\gamma(\hat{\theta}_0)) - 4\nabla\gamma(X)^T(\hat{\theta}_0 - \hat{\theta}_K) \leq 0. \tag{2.27}$$

2.4.3 Example

Choose $\gamma(\hat{\theta}_0) = c/l^2$ where c is an arbitrary constant and $l^2 = \|\hat{\theta}_0 - \hat{\theta}_K\|^2_\Sigma$. Let Σ be a completely specified covariance matrix. Further, let J denote the Jacobian of the projection of the least squares estimator onto the Lasso constraint set. From Lemma 1.6 we get that the Jacobian J is

$$J = \frac{\partial\hat{\theta}_K}{\partial\hat{\theta}_0} = (Z^TZ)^{-1}Z^TH_\lambda(y)Z \tag{2.28}$$

where $H_\lambda(y)$ is the hat matrix for any value of the Lasso tuning parameter λ, on the subspace spanned by the covariates corresponding to the set of nonzero coefficients. In other words, for any arbitrary λ, let $A = \{i : \hat{\theta}_i \neq 0\}$. Then A is called the active set and $H_\lambda(y) = Z_A(Z_A^TZ_A)^{-1}Z_A^T$ where Z_A is the set of covariates corresponding to the active set A. It can be shown that,

$$\nabla\gamma = \frac{-2c}{l^4}[\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K) - J\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K)],$$

and

$$\Sigma\nabla\gamma = -\frac{2c}{l^4}[(\hat{\theta}_0 - \hat{\theta}_K) - \Sigma J\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K)].$$

Calculating the divergence on both sides yields,

$$\begin{aligned}\text{div}\{\Sigma\nabla\gamma\} = -2c[(\nabla\frac{1}{l^4})^T(\hat{\theta}_0 - \hat{\theta}_K) + \frac{1}{l^4}\text{div}(\hat{\theta}_0 - \hat{\theta}_K) \\ - (\nabla\frac{1}{l^4})^T\Gamma(\hat{\theta}_0 - \hat{\theta}_K) - \frac{1}{l^4}\text{div}\{\Gamma(\hat{\theta}_0 - \hat{\theta}_K)\}],\end{aligned} \tag{2.29}$$

where $\Gamma = \Sigma J \Sigma^{-1}$. Similarly,

$$\nabla(\frac{1}{l^4}) = -\frac{4}{l^6}[\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K) - J\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K)],$$

$$\begin{aligned}\nabla(\frac{1}{l^4})^T(\hat{\theta}_0 - \hat{\theta}_K) &= -\frac{4}{l^6}[l^2 - (\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T(\hat{\theta}_0 - \hat{\theta}_K)] \\ &= -\frac{4}{l^4} + \frac{4}{l^6}(\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T(\hat{\theta}_0 - \hat{\theta}_K),\end{aligned}$$

and

$$\begin{aligned}\nabla(\frac{1}{l^4})^T\Gamma(\hat{\theta}_0 - \hat{\theta}_K) = -\frac{4}{l^6}[&(\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}\Gamma(\hat{\theta}_0 - \hat{\theta}_K) \\ &- (\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T\Gamma(\hat{\theta}_0 - \hat{\theta}_K)] \\ = -\frac{4}{l^6}[&(\hat{\theta}_0 - \hat{\theta}_K)^T J\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K) \\ &- (\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T\Gamma(\hat{\theta}_0 - \hat{\theta}_K)].\end{aligned}$$

Since $\mathrm{div}(\hat{\theta}_0 - \hat{\theta}_K) = m$, it can be shown that

$$\begin{aligned}\mathrm{div}(\Gamma(\hat{\theta}_0 - \hat{\theta}_K)) &= \mathrm{tr}(\Gamma) - \mathrm{tr}(\Gamma J^T) \\ &= \mathrm{tr}(J) - \mathrm{tr}(\Gamma J^T) \\ &= p - m - \mathrm{tr}(\Gamma J^T).\end{aligned}$$

Hence we can write (2.29) as

$$\begin{aligned}\mathrm{div}\{\Sigma\nabla\gamma\} = -\frac{2c}{l^4}[&(m-4) + \frac{4}{l^2}(\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T(\hat{\theta}_0 - \hat{\theta}_K) \\ &+ \frac{4}{l^2}(\hat{\theta}_0 - \hat{\theta}_K)^T J\Sigma^{-1}(\hat{\theta}_0 - \hat{\theta}_K) \\ &- \frac{4}{l^2}(\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T\Gamma(\hat{\theta}_0 - \hat{\theta}_K) \\ &- (p - m - \mathrm{tr}(\Gamma J^T)]\end{aligned} \tag{2.30}$$

and

$$(\nabla\gamma)^T(\hat{\theta}_0 - \hat{\theta}_K) = -\frac{2c}{l^4}[l^2 - (\hat{\theta}_0 - \hat{\theta}_K)^T\Sigma^{-1}J^T(\hat{\theta}_0 - \hat{\theta}_K)]. \tag{2.31}$$

Therefore, substituting (2.30) and (2.31) in (2.27) we get a quadratic in c from which considering $c > 0$ we get a sufficient condition for domination over an unbiased loss estimator when Lasso is used as an estimator for the linear model. In Section 2.5 we present a detailed simulation study using this choice of $\gamma(\hat{\theta}_0)$ for varying levels of sparsity in the "true" model.

2.5 Simulation

We consider the classical linear model framework $Y = Z\theta + \epsilon$ where Z is a $n \times p$ matrix of covariates with full column rank p, and $\epsilon \sim N(0, \sigma^2 I_n)$. The simulation is conducted for varying levels of sparsity in the unknown vector of regression coefficients θ. Here, we present the results when $n = 1000$ and $p = 100$ but the results look similar for other settings of n and p under the condition that $p < n$.

The results in Table 2.2 show the risk gains of the loss estimate itself obtained on using the improved loss estimator when Lasso is used as the estimator of θ for varying levels of sparsity. By sparsity or zeroes in the model, we mean the "true" number of regression coefficients, θ, that are zero. The first column in Table 2.2 is denoted by A, the active set of nonzero parameters. The parameter vector, $\theta \in \mathbb{R}^{100}$ is a 100 dimensional vector. The active set A shows the "true" number of non-zero components in the θ vector. The active set decreases from 90 to 10 where 90 indicates that the true number of zeroes in the model are 10. Similarly, $A = 10$ indicates that the true number of zeroes in the linear model is 90. The next five columns denoted by RG indicates the risk gains obtained on using the improved loss estimator under different settings of the error variance, σ^2. The table shows risk gains for different sparsity levels as the error variance increases from $\sigma^2 = 1$ to $\sigma^2 = 6000$. The simulation methodology is outlined in the next paragraph.

Let RD equal the left hand side in (2.27). We obtain the percent risk gains as $RG = -100 * RD/\lambda^U$. Generate a p dimensional θ^* vector with A number of non-zero components in θ^*. Generate an $n \times p$ matrix Z, from a multivariate normal distribution such that the correlation between Z_i and Z_j is $\rho = 0.5^{|i-j|}$. Setting $y_i = \sum_{j=1}^{p} z_{ij}\theta_j^* + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, we generate the response y. Performing the LARS algorithm on the generated dataset, we get the entire LASSO solution path and pick the optimal Lasso solution as the one for which C_p is the lowest. Let $\hat{\theta}_0$ and $\hat{\theta}_K$ denote the least squares and the optimal Lasso solution respectively. Choosing $\gamma(\hat{\theta}_0) = c/l^2$, we obtain an upper bound for the sufficient condition for domination of $\lambda^g(\hat{\theta}_0) = \lambda^U(\hat{\theta}_0) - \gamma(\hat{\theta}_0)$ as given in (2.27). Denote that upper bound by C. Note that the interval, $(0, C)$ is a *random* interval because it depends on the Lasso solution. Therefore, for any $0 < c < C$, we improve in estimating the loss if Lasso is used as an estimator. We report our simulation results by choosing $c = C/2$. This is admittedly a sub-optimal choice but theoretical optima of c is as yet unclear. We carry out 5000 simulation runs for each setting of the active set A and error variance σ^2 combination.

Based on this choice of c, we note from Table 2.2 that we observe a similar pattern of risk gains. We observe maximal risk gains of around $7\% - 8\%$ in the middle ranges, i.e., when A/p is between 0.30 and 0.70. The risk gains are in the order of 4% when the true number of zeroes in the model is large, i.e., when $A/p \to 0$. It is encouraging to see that we get significant risk gains of around $4\% - 7\%$ even when the number of zeroes in the true model is few as evidenced by the second row in Table 2.2 . Further, as the error variance increases, which is a realisitic situation, we observe larger gains in risk on using the improved estimator for loss. This fact can be gleaned as we move across from left to right for any given row

in the table. In particular, when the error variance is really large, we observe risk gains of around 10% when $A/p \to 1$. Consequently, we note that in any modelling situation where Lasso is used, we could obtain risk gains of around $5\% - 9\%$ depending on the number of zeroes as estimated by Lasso. An interesting curious fact observed in the simulations is the

Table 2.2 – Percentage Risk Gains for different Active Sets for $n = 1000$ and $p = 100$

A	RG ($\sigma^2 = 1$)	RG ($\sigma^2 = 80$)	RG ($\sigma^2 = 300$)	RG ($\sigma^2 = 2000$)	RG ($\sigma^2 = 6000$)
90	1.13	1.86	2.78	6.79	10.55
80	4.09	4.65	5.69	7.14	9.17
70	6.96	7.11	7.52	8.16	7.95
60	7.85	7.78	7.93	7.89	6.56
50	7.90	7.01	7.13	6.10	4.46
40	5.97	5.98	5.09	5.29	4.66
30	5.15	5.60	5.22	4.62	3.50
20	4.70	4.51	4.66	3.81	3.60
10	4.27	4.31	3.01	3.68	7.42

distribution of the squared length of projection $l^2 = \|\hat{\theta}_0 - \hat{\theta}_K\|^2_\Sigma$ for varying levels of sparsity. For each active set (A) and error variance (σ^2) combination, we present the QQ-plots in Figures 2.1, 2.2, 2.3 and 2.4 respectively. Each figure shows the QQ-plots of the squared length of projection, l^2, for varying active sets, ranging from $A = 90$ to $A = 10$ overlaid with a fitted Weibull distribution for error variances ranging from $\sigma^2 = 1$ to $\sigma^2 = 6000$. The Weibull parameters are estimated using their corresponding maximum likelihood estimators.

In Figure 2.1 for instance, the QQ-plots for active sets ranging from $A = 90$ to $A = 10$ are presented when the error variance in the linear model is 1. Figure 2.2 presents a similar picture when the error variance is 300.

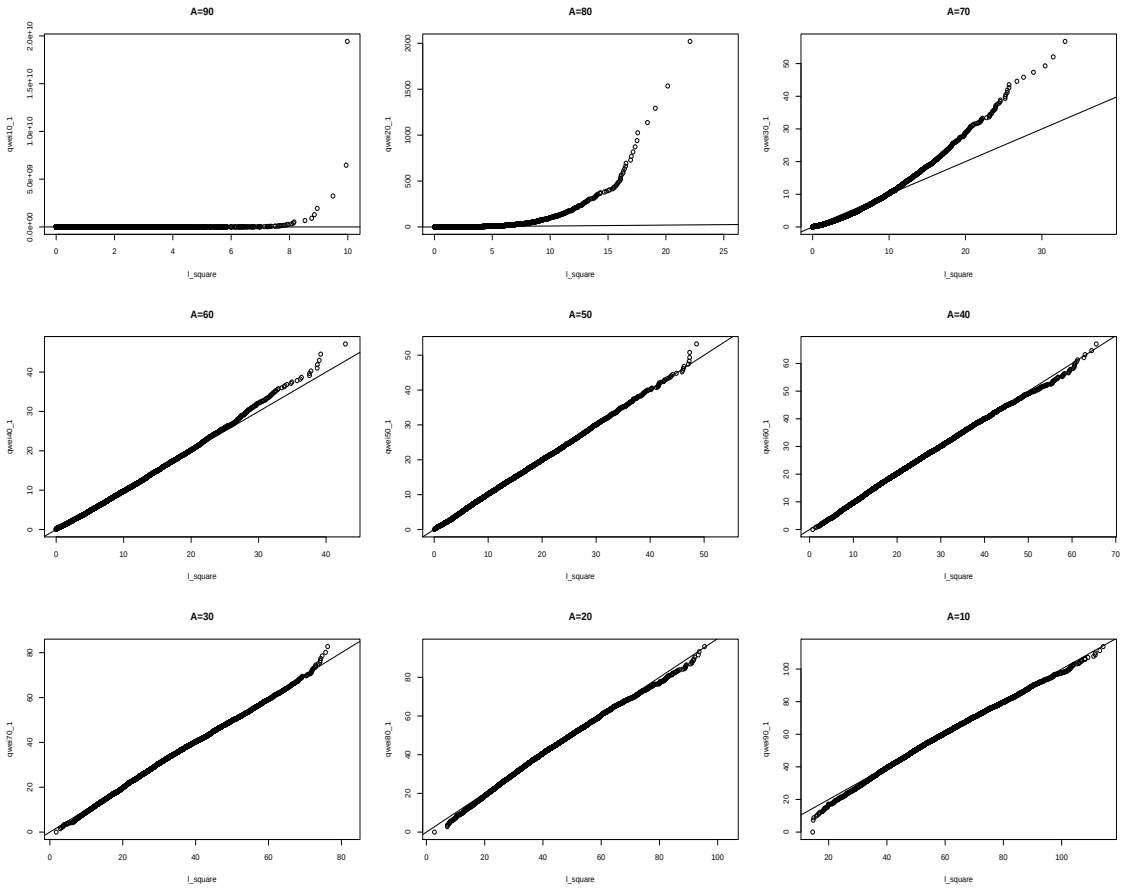

Figure 2.1 – QQ plot of l^2 with Weibull Distribution for $n = 1000$, $p = 100$, $\sigma^2 = 1$

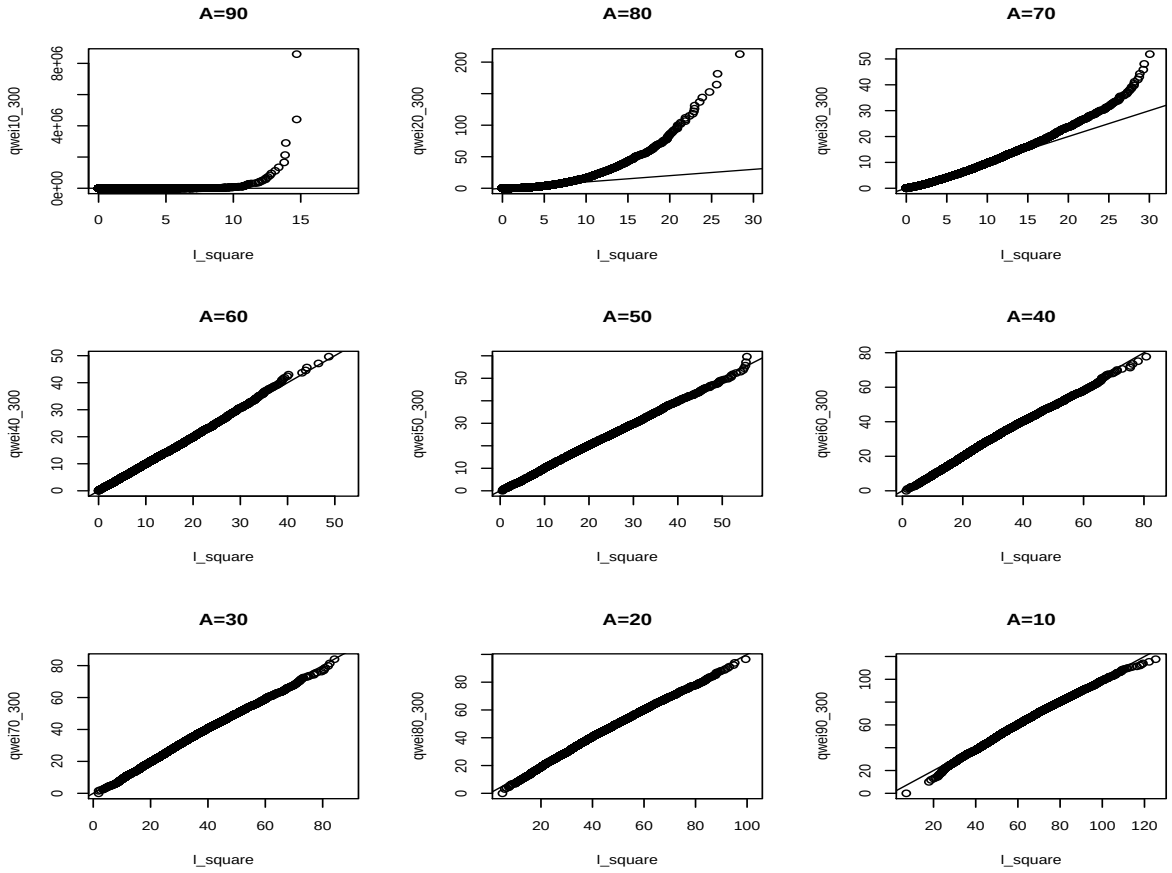

Figure 2.2 – QQ plot of l^2 with Weibull Distribution for $n = 1000$, $p = 100$, $\sigma^2 = 300$

It seems from the QQ-plots that irrespective of the error variance in the "true" model, Weibull distribution seems to fit very well when the active sets exceed some threshold. From the figures shown here, it seems that the fit is quite good when the active set exceeds 70. We observe that the distributions of l^2 change from being a quasi-exponential distribution to a quasi-Gaussian distribution as A changes from 90 to 10. In other words, when $A = 90$, the distribution of l^2 looks like an exponential distribution and when A is small, i.e., $A = 10$, the distribution of l^2 looks like a Gaussian distribution. For intermediate values of A, the distribution seems right skewed, akin to a gamma distribution. This flexible shape of the variable

based on the extent of "true" sparsity in the model prompted to fit Weibull distribution to the data. As a special case, observe that if in the constraint set $K_t := \{\theta : \sum_{j=1}^{p} |\theta_j| \leq t\}$, we set $t = 0$, then $\hat{\theta}_K = 0$ and as such $l^2 = \|\hat{\theta}_0\|_{\Sigma}^2$, becomes the norm of the least squares vector which we know has a χ^2 distribution.

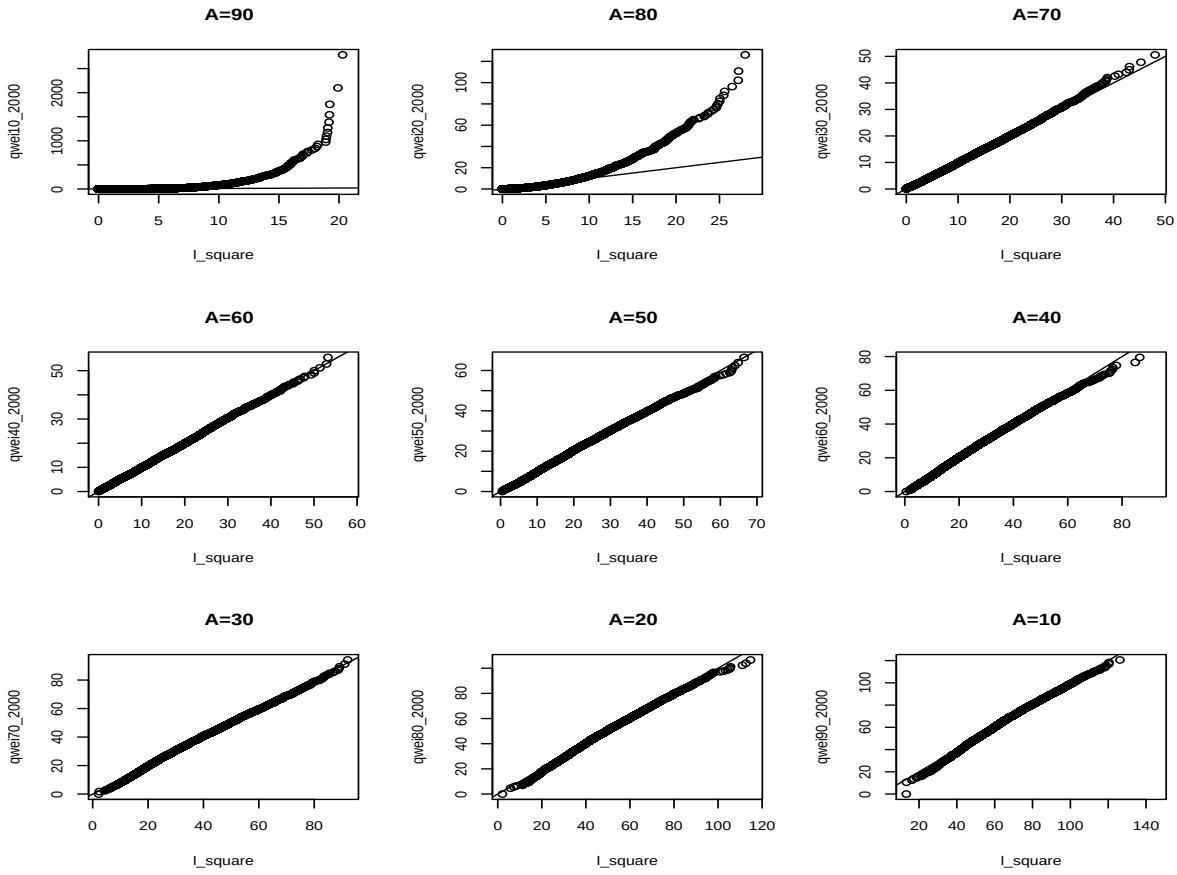

Figure 2.3 – QQ plot of l^2 with Weibull Distribution for $n = 1000$, $p = 100$, $\sigma^2 = 2000$

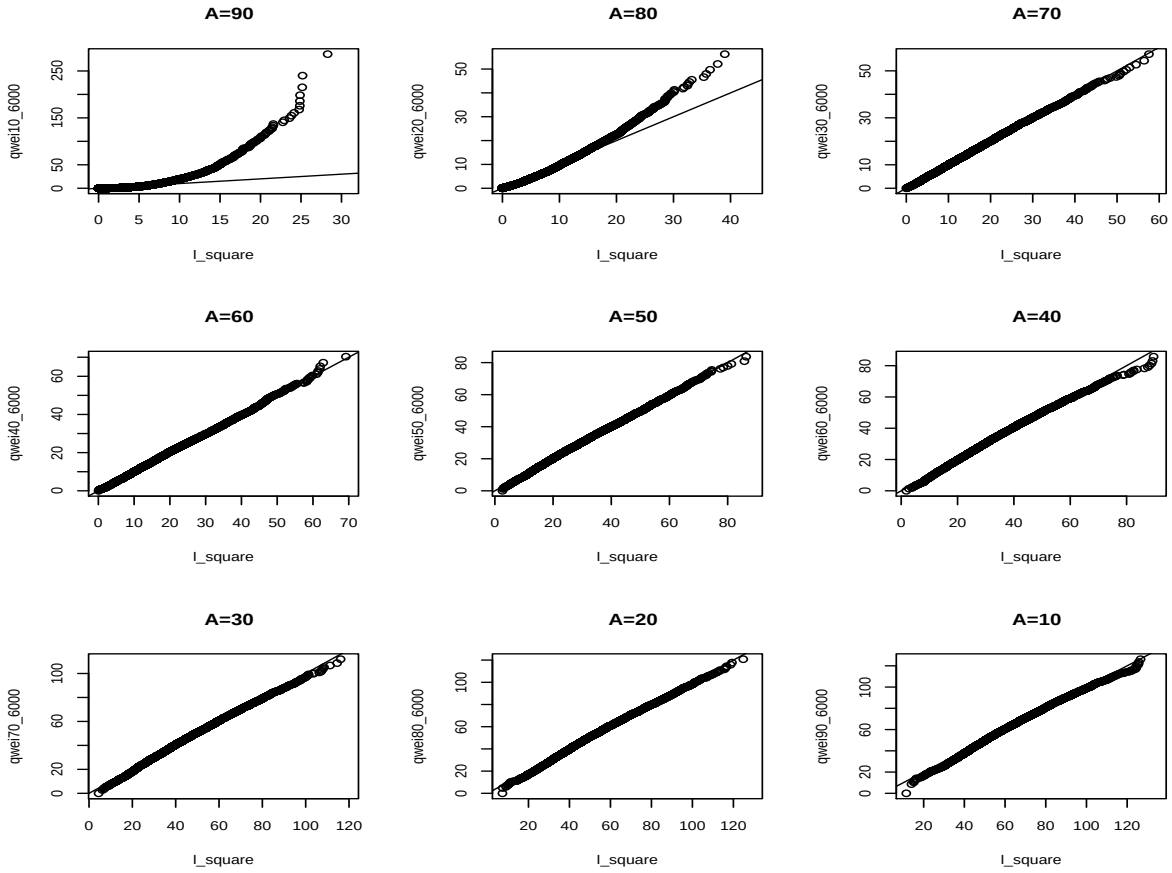

Figure 2.4 – QQ plot of l^2 with Weibull Distribution for $n = 1000$, $p = 100$, $\sigma^2 = 6000$

The fitted Weibull distribution to the empirical data of l^2 is presented in Table 2.3. The first column gives the error variance of the model. The second column shows the fitted shape and scale parameters of the Weibull distribution. The next 9 columns present the values of the fitted shape and scale parameters for the squared length projection, l^2, as the active set decreases from $A = 90$ to $A = 10$. We observe that for a given error variance, the shape parameter increases steadily from less than 1 to about 4 and the corresponding scale parameters increase steeply from small positive values till about 70 as A decreases from 90 to 10. It is also discernible that the estimated shape and scale parameters increase as the

error variance increases for a fixed value of A. For instance, when $A = 90$, proceeding down along that column, one notices that the fitted shape and scale parameters for $\sigma^2 = 1$ are 0.08 and 0.02 respectively and the fitted shape and scale parameters for $A = 90$ and $\sigma^2 = 6000$ are 0.49 and 3.49 respectively. We can thus conjecture that the fitted parameter estimates seem to depend on the error variance of the true model.

Table 2.3 – Weibull fits of l^2 for different error variance and active sets

Variance	Weibull	Active Sets from 90 to 10								
$\sigma^2 = 1$	shape	0.08	0.31	0.98	1.50	1.83	2.33	2.67	3.02	3.90
	scale	0.02	1.96	6.32	11.33	16.50	26.74	37.10	47.22	65.61
$\sigma^2 = 80$	shape	0.11	0.33	0.90	1.67	2.04	2.27	2.57	3.16	3.78
	scale	0.20	2.27	6.42	13.77	20.26	27.43	34.71	50.15	63.84
$\sigma^2 = 300$	shape	0.13	0.54	1.12	1.64	1.99	2.47	2.68	3.09	4.33
	scale	0.42	3.91	7.70	13.43	20.36	32.68	37.72	48.37	71.76
$\sigma^2 = 2000$	shape	0.30	0.66	1.51	1.70	2.16	2.23	2.67	3.40	4.39
	scale	2.13	4.86	12.19	15.75	24.62	30.38	42.19	56.73	74.01
$\sigma^2 = 6000$	shape	0.49	1.11	1.69	1.85	2.49	2.59	2.88	3.30	3.91
	scale	3.49	8.23	16.11	22.09	35.42	37.48	53.22	63.13	73.69

2.6 Discussion and Future Work

We have exhibited improved estimators of the loss function for the case of estimating the mean vector in the multivariate normal model with an arbitrary covariance matrix when the mean vector is estimated using (1) MLE and (2) an improved estimator. Extending

those ideas, we presented improved estimators of the loss difference when the competing estimators of the parameters of a linear model are the least squares estimator and Lasso respectively. In addition, we derived sufficient conditions under which domination over an unbiased estimator of loss can be obtained and presented simulation studies to demonstrate the theoretical results. The optimal choice of c in $\gamma(\hat{\theta}_0)$ for the loss estimation calculations in the linear model setting is not clear. Investigating methods to choose an optimal c in the random interval would certainly result in larger risk gains than demonstrated here. As far as improved loss estimators are concerned, extensions of the techniques developed here could also be used to devise improved loss estimators for more general constrained regression problems like group Lasso and fused Lasso.

The simulation study in the linear model framework was particularly illuminating in the context of the distribution of the squared length of projection. We conjecture that if $\hat{\theta}_0$ and $\hat{\theta}_K$ denote the least squares and Lasso estimators of the regression coefficients of a linear model and suppose $l^2 = \|\hat{\theta}_0 - \hat{\theta}_K\|^2_\Sigma$ denote the squared distance between the two estimators and further suppose k denotes the "true" number of zeroes in the linear model, then, $l^2 \sim W(\xi, \lambda)$, has a Weibull distribution with shape parameter ξ and scale parameter λ as $k(n)/p(n) \to \nu \in (\alpha, 1)$, where n is the number of observations and α is a positive real number bounded away from zero.

The conjecture has more general ramifications in distribution theory where one could think of deriving distributions of lengths of projections of a random vector with a correlation structure when the parameter space lies in more general constraint sets like a cone or a convex set with a smooth boundary. Kuriki and Takemura [30] derived the conditional distribution of

the squared length of projection for a multivariate normal random vector in the i.i.d. setting when the parameter space was a closed convex set with a piecewise smooth boundary. It would be instructive to extend the idea to a multivariate normal random vector with a covariance structure and then to explore the distributional properties of squared lengths for spherically symmetric and ellipitically symmetric distributions.

Chapter 3

Largest Eigenvalue Distribution

3.1 Introduction

Eigenvalues of random matrices have a rich mathematical structure and are a source of interesting distributions. The distribution theory for a large number of multivariate statistical procedures, such as principle component, canonical correlation, discriminant analysis, multivariate analysis of variance (MANOVA), Roy's union intersection test, and so forth, are derived from the extreme eigenvalues of certain random matrices. Here, we present a unified approach to find the exact distribution of the largest eigenvalue in the Gaussian and the double Wishart settings, the latter being ubiquitous in the theory of multivariate statistical analysis.

For square symmetric random matrices, the celebrated semicircle law of Wigner [54] describes the limiting density of eigenvalues; the analogue for covariance matrices is due to Marčenko and Pastur [32]. Beginning in the 1950s, physicists began to use random matrix models to study quantum phenomena. Many early results in random matrix theory (RMT) have their origins in quantum mechanics where the energy levels of a system are described using the eigenvalues of a Hermitian operator H (known as the Hamiltonian). Wigner proposed that the local statistical behaviour of energy levels can be well modelled using the eigenvalues of a large random matrix. In particular, Wigner [54] gives the famous semicircular law which states that the empirical spectral density of an $n \times n$ symmetric random matrix with i.i.d.

entries having mean 0 and variance σ^2 has a semicircular law given by

$$dF(x\sigma\sqrt{n}) = \frac{1}{4\pi}\sqrt{4-x^2}dx.$$

for $-2 < x < 2$.

Even for Gaussian random matrices, the exact distributions of the largest roots are difficult to compute. Most exact expressions are in terms of a hypergeometric function with a matrix argument, with no general and simple closed form. Recently some exact algorithms have been made available, but they are not yet in wide use. Koev and Edelman [29] have developed efficient algorithms (and a MATLAB package available at `http://www-math.mit.edu/~plamen`) for the evaluation of such matrix hypergeometric functions using recursion formulae from group representation theory.

The elegant approximations in Johnstone [24, 25] to the distribution of the largest roots of Wishart random matrix and the multivariate beta distribution turn out to be expressed in terms of the Tracy-Widom distributions from RMT. These distributions are parameter free, and can be easily tabulated. The approximation is an asymptotic one, in which the dimension p increases to infinity, and the degrees of freedom parameters grow in proportion to p. It can be shown that these approximations are of the order $O((p \wedge n)^{-1/3})$. We bypass these approximations and present an exact distribution theory for the distribution of largest roots using a class of determinant formulae and a remarkable theory of integration for determinants developed by De Bruijn [6].

3.1.1 The Wishart distribution

Recall that if X_j are i.i.d. standard normal variables, $N(0,1)$, then the distribution of $\chi^2_n = X_1^2 + X_2^2 + \ldots + X_n^2$ has the central χ^2 distribution. When measurements are made

on p continuous variables on n independent subjects, it is common in multivariate statistics to assume the data to be n i.i.d. copies of a p-variate normal distribution with an unknown mean vector and a fixed but unknown covariance matrix, i.e., for a $n \times p$ random matrix $X = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}]$ and each row $\mathbf{X_j} \sim N_p(\boldsymbol{\mu}, \Sigma)$. We can thus say that $\mathbf{X} \sim N_{np}(\mathbf{1} \otimes \boldsymbol{\mu}, \mathbf{I_n} \otimes \boldsymbol{\Sigma})$. If $A = X^T X$, where the $n \times p$ matrix X is $N(0, I_n \otimes \Sigma)$, then A is said to have the *Wishart distribution* with n degrees of freedom and covariance matrix Σ. We will write A is $W_p(n, \Sigma)$. The Wishart distribution plays an important role in multivariate analysis because given a data matrix, $X^{n \times p}$ regarded as n i.i.d. copies of a multivariate normal distribution, the sample covariance matrix defined as

$$S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T$$

is $W_p(n, \frac{\Sigma}{n})$. If A is $W_p(n, \Sigma)$ with $n \geq p$, then the density function of A is given by

$$f(A) = \frac{1}{2^{\frac{pn}{2}} \Gamma_p(\frac{n}{2}) (\det \Sigma)^{\frac{n}{2}}} \text{etr}(-\frac{1}{2}\Sigma^{-1}A)(\det A)^{\frac{n-p-1}{2}} \tag{3.1}$$

where $\Gamma_p(\frac{n}{2})$ denotes the multivariate gamma function defined by

$$\Gamma_p(a) = \int_{A>0} \text{etr}(-A)(\det A)^{a-\frac{p+1}{2}}(dA) \tag{3.2}$$

for $[Re(a) > (p-1)/2]$ and $\text{etr}(A) = \exp\{\text{tr}(A)\}$ and the integral is over the space of $p \times p$ positive definite matrices (note that for $p = 1$ in (3.2), the right hand side reduces to the univariate Gamma function). The density function of the central χ^2_n can be obtained as a special case by substituting $p = 1$ and $\Sigma = 1$.

Many multivariate analysis techniques like principal components, MANOVA, canonical correlations depend on the eigen-analysis of sample covariance matrices. The characteristic roots of a nonsingular A are real and nonnegative, and according to the classical formula of

Fisher, Girshick, Hsu, Mood and Roy, have a density which is given in Muirhead [34, p. 106] by

$$f(l_1,\ldots,l_p) = C(p,n)\prod_{i=1}^{p} l_i^{\frac{n-p-1}{2}} \prod_{i<j}^{p}(l_i - l_j) \int_{\mathcal{O}(p)} \text{etr}(-\frac{1}{2}\Sigma^{-1}HLH')dH, \tag{3.3}$$

where $C(p,n)$ is a normalising constant that can be explicitly evaluated, $L = \text{diag}(l_1, l_2, \ldots, l_p)$, $H \in \mathcal{O}(p)$ is a an orthogonal matrix belonging to $\mathcal{O}(p)$, the orthogonal group of $(p \times p)$ matrices, dH represents the Haar invariant measure on $\mathcal{O}(p)$ normalised so that the volume of $\mathcal{O}(p)$ is one.

An explicit closed form formula for the above density function is difficult to calculate owing to the integral over the orthogonal group. However, infinite series expansions for the integral are available using the theory of *Zonal Polynomials*, a detailed treatment of which can again be found in Muirhead [34]. Constantine [9] expresses the cumulative distribution function of the largest root distribution in terms of a matrix hypergeometric function as

$$P(l_{max} < nx) = d_{p,n}x^{\frac{pn}{2}}{}_1F_1(\frac{n}{2}, \frac{n+p+1}{2}; -\frac{n}{2}xI_p), \tag{3.4}$$

where $d_{p,n}$ is a constant depending only on p and n (cf. [34, pg.421]).

For a wide range of modern data sets (microarray data, genomics, weather forecasting, etc.), the number of features p is very large while the number of observations n is much smaller than or just comparable to p. For these situations, the classical asymptotics is not always appropriate and different asymptotic theories are needed. When $n < p$, and A is singular and Srivastava [42] gives the probability density function for the singular Wishart case. We are however interested in the classical $n > p$ case. In the next subsection, we briefly discuss the role of eigenvalues in the double Wishart setting.

3.1.2 Double Wishart Setting

Let A and B be independent, central Wishart matrices in p variables with common covariance and having n_1 and n_2 degrees of freedom respectively. Let Σ be the covariance matrix of the multivariate normal distribution from which the Wishart matrices have been constructed. Define, $C = (A + B)^{-1}B$. We are interested in computing the distribution of the largest eigenvalue of C. This problem is equivalent to solving the determinantal equation $\det[B - l(A + B)] = 0$. In some hypothesis testing situations, interest may be in constructing the likelihood ratio test which would require the joint distribution of all the eigenvalues but in many other situations, the inference is based on looking at the distribution of the largest characteristic root.

Definition 3.1. (Real, Complex and Symplectic Jacobi Ensembles) *Let $A \sim W_p^{(\beta)}(n_1, I)$ be independent of $B \sim W_p^{(\beta)}(n_2, I)$ where $n_1, n_2 \geq p$ are independent real, complex or quaternion Wishart matrices (where $\beta = 1, 2$ or 4 respectively), then $C = (A+B)^{-1}B$ is called the real, complex or symplectic Jacobi ensemble indicated by the parameter β. In particular, the real Jacobi ensemble is called the multivariate beta distribution.*

Definition 3.2. (Greatest root statistic) *The largest eigenvalue l_1 of $C = (A + B)^{-1}B$, the multivariate beta distribution is called the greatest root statistic and a random variable having this distribution is denoted by $l_1(p, n_1, n_2)$.*

Since A and B are positive definite, all the eigenvalues are less than 1. Moreover, every eigenvalue of the real, complex or symplectic Jacobi ensemble is real. Let $l_1, l_2, \ldots, l_p$ denote the eigenvalues of C. The joint density of the eigenvalues is given by

$$f_\beta(\mathbf{l}) = \frac{1}{I(p, \beta, a_1, a_2)} \prod_{i=1}^{p} l_i^{a_1 - \frac{\beta(p-1)}{2} - 1} (1 - l_i)^{a_2 - \frac{\beta(p-1)}{2} - 1} \prod_{i<j} |l_i - l_j|^\beta, \tag{3.5}$$

where

$$I(p,\beta,a_1,a_2) = \frac{\Gamma_p^{(\beta)}(1+\frac{\beta}{2}p)}{\pi^{\frac{\beta p(p-1)}{2}}(\Gamma(1+\frac{\beta}{2}))^p}\frac{\Gamma_p^{(\beta)}(a_1)\Gamma_p^{(\beta)}(a_2)}{\Gamma_p^{(\beta)}(a_1+a_2)}.$$

is the Selberg Integral value [39] and

$$\Gamma_p^{(\beta}(c) = \pi^{\frac{p(p-1)\beta}{4}}\prod_{i=1}^{p}\Gamma(c-\frac{\beta}{2}(i-1)).$$

for $Re(c) > \frac{\beta}{2}(p-1)$ is the multivariate gamma function with parameter $\beta > 0$. While $\beta = 2, 4$ are interesting mathematical objects, the case corresponding to $\beta = 1$, the real case, is more relevant for the purposes of multivariate statistical analysis and hence for the most part of this chapter, we focus on the case $\beta = 1$. In (3.5), it is known that $a_1 = n1/2$ and $a_2 = n_2/2$. As such, these parameters are known in advance and need not be estimated using any statistical procedure. Exact evaluation of the marginal distribution of the largest eigenvalue has not been found yet and in this note, we propose a method for the same. Dumitriu et al. in [15], give an expression for the distribution function and the density function of the largest eigenvalue of a real Jacobi matrix as

$$P(l_1 < x) = C_{1,p}x^{\frac{pn_1}{2}}{}_2F_1(\frac{n_1}{2}, \frac{-n_2+p+1}{2}; \frac{n_1+p+1}{2}; xI), \tag{3.6}$$

where

$$C_{1,p} = \frac{\Gamma_p^{(1)}(\frac{n_1+n_2}{2})\Gamma_p^{(1)}(\frac{p+1}{2})}{\Gamma_p^{(1)}(\frac{n_1+p+1}{2})\Gamma_p^{(1)}(\frac{n_2}{2})}.$$

and ${}_2F_1(\cdot,\cdot;\cdot,xI)$ denotes the hypergeometric function with a matrix argument, which in this case is considered to be the identity matrix. They further provide a generalisation of the above result for any $\beta > 0$ and not just the real case.

3.1.3 Multiple Integrals and Determinants

In this subsection, we review a classical result involving multiple integrals and determinants that are used in subsequent sections to compute some multiple integrals having determinantal

representations. We present it here for the sake of completeness. The following result was established in the late nineteenth century.

$$\int_{a<x_1\leq\ldots\leq x_n<b}\cdots\int \det(\phi_i(x_j))\det(\psi_i(x_j))dx_1\ldots dx_n = \det_{1\leq i,j\leq n}\int_a^b \phi_i(x)\psi_j(x)dx. \tag{3.7}$$

A formula for a similar integral with a single determinant term was not established until 1955 by De Bruijn in the seminal paper [6]. He provided a technique to compute integrals of the following form in an ordered measure space.

$$\Omega = \int_{a<x_1\leq\ldots\leq x_n<b}\cdots\int \det\phi_i(x_j)dx_1dx_2\ldots dx_n. \tag{3.8}$$

The result for this type of integral was similar to the previous one, except that instead of Ω, its square was expressed as a determinant; in other words Ω was expressed as a *Pfaffian* form.

The computation of the above integral is closely related to the *s*ignature function, $E(x_1, x_2, \ldots, x_n)$ defined as follows. For any ordered set S, if $x_1 \in S, x_2 \in S, \ldots, x_n \in S$, the signature function satisfies two properties:

(1) $E(x_1, x_2, \ldots, x_n) = 1$ if $x_1 < x_2 < \ldots < x_n$ and (2) $E(x_1, x_2, \ldots, x_n)$ is alternating in its arguments. Consequently, we can write it as

$$E(x_1, x_2, \ldots, x_n) = \prod_{1\leq i<j\leq n}\text{sgn}(x_j - x_i).$$

Note that the range of integration is not restricted to finite values of a and b. It is true even if $a = -\infty$ and $b = \infty$. The approach holds true for any ordered measure space. The following expansion of the signature function can be proved by the method of induction. Considering n to be even and $m = n/2$, we can write

$$E(x_1, x_2, \ldots, x_n) = \frac{1}{2^m m!}\sum_{j_1=1}^{n}\cdots\sum_{j_n=1}^{n} E(j_1, \ldots, j_n)E(x_{j1}, x_{j2}), \ldots$$
$$\ldots E(x_{j_{2m-1}}, x_{j_{2m}})$$

where $E(j_1, \ldots, j_n)$ is the signature of the permutation of $j_1, \ldots, j_n$ and is equal to zero when it is not a permutation. The formulaic structure of the signature function is similar to the representation of a *Pfaffian* which is defined as follows.

Definition 3.3. *Given a skew symmetric $n \times n$ matrix $A = (a_{ij})$, the Pfaffian of A is given by*

$$Pf(A) = \frac{1}{2^m m!} \sum_{j_1=1}^{n} \ldots \sum_{j_n=1}^{n} E(j_1, \ldots, j_n) a_{j1} a_{j2} \ldots a_{j_{2m-1}} a_{j_{2m}},$$

where we let n to be even and $m = n/2$. It is equivalently given by

$$= \frac{1}{2^m m!} \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{m} a_{\sigma(2i-1),\sigma(2i)},$$

where S_n is the symmetric group and $sgn(\sigma)$ is the signature of the permutation σ.

Note that multiplying the integrand on the right hand side of (3.8) by $E(x_1, x_2, \ldots, x_n)$ leaves the value of the integral unaltered and makes the integrand symmetric in its argument. Consequently, we can write (3.8) as

$$\Omega = \frac{1}{n!} \int_a^b \ldots \int_a^b E(x_1, \ldots, x_n) \det \phi_i(x_j) dx_1 \ldots dx_n.$$

Using symmetry in the arguments of signature function and the fact that the determinant of an $n \times n$ matrix has $n!$ terms we can rewrite the integral as

$$\Omega = \int_a^b \ldots \int_a^b E(x_1, \ldots, x_n) \phi_1(x_1) \ldots \phi_n(x_n) dx_1 \ldots dx_n.$$

Noting the expansion of the signature function from its definition, the integral Ω can be written as a sum of products of m double integrals if n is even and an additional double integral if n is odd and we thus get the following theorem.

Theorem 3.1. (DeBruijn, 1955)

$$(i) \quad \int \cdots \int \det(\phi_j(x_k))_{1\le j,k\le N} \det(\psi_j(x_k))_{1\le j,k\le N} d\mu(x_1)\ldots d\mu(x_N)$$
$$= N! \det\left(\int \phi_j(x)\psi_k(x)\right)_{1\le j,k\le N}; \quad (3.9)$$

$$(ii) \quad \int\limits_{a<x_1\le\ldots\le x_N<b} \cdots \int \det(\phi_j(x_k)) d\mu(x_1)\ldots d\mu(x_N)$$
$$= Pf\left(\int_a^b \int_a^b \operatorname{sgn}(x-y)\phi_j(x)\phi_k(y)d\mu(x)d\mu(y)_{1\le j,k\le N}\right); \quad (3.10)$$

$$(iii) \quad \int \cdots \int \det(\phi_j(x_k) \quad \psi_j(x_k))_{1\le j\le 2N,1\le k\le N} d\mu(x_1)\ldots d\mu(x_N)$$
$$= (2N)! Pf\left(\int \phi_j(x)\psi_k(x) - \phi_k(x)\psi_j(x)d\mu(x)\right)_{1\le j,k\le 2N}. \quad (3.11)$$

These integral identities were developed in de Bruijn [6] as an attempt to generalise (3.7). Note that the first and last integral identities are valid in general measure spaces. In the second identity, the space needs to be ordered. In the last identity, the left hand side determinant is a $2N \times 2N$ determinant whose columns are alternating columns of the ϕ_j and ψ_j, hence the notation, and asymmetry in indexing. It is quite interesting that most of the foundational theory of random matrices, in the case of invariant measures, is a consequence of the integrals given in Theorem 3.1.

3.2 Results for Gaussian Ensembles

We first consider the case of real valued Gaussian matrices which will be referred to as Gaussian Orthogonal Ensemble corresponding to the case of $\beta = 1$. The complex and the quaternion cases are simple extensions of the real case.

Definition 3.4. *Consider matrices of the form,* $M_n = [X_{i,j}]_{i,j=1}^n$ *where* $X_{i,j} = X_{j,i} \sim N(0,1)$, $i < j$ *and* $X_{i,i} \sim \sqrt{2}N(0,1)$ *and are independent for all* (i,j). *Such a matrix is called a Gaussian Orthogonal Ensemble.*

The above random matrix can be constructed in the following way. Let $A = [Y_{i,j}]_{i,j=1}^n$, where $Y_{i,j} \sim N(0,1)$ where the entries are independent and identically distributed. Defining $M_n = (A + A^T)/\sqrt{2}$ would give the required Gaussian matrix.

Let $\mathbb{H}_n$ denote the space of $(n \times n)$ real symmetric matrices, which has $\binom{n}{2} + n$ free variables. Thus a canonical measure on this space is the Lebesgue measure on $\mathbb{R}^{\binom{n}{2}+n}$. The density for the GOE with the above Lebesgue measure is $\frac{1}{Z_n} \exp[-\frac{1}{4}\mathrm{tr}H^2]dH$ where

$$\begin{aligned}\mathrm{tr}H^2 &= \sum_{i,j=1}^n h_{ij}^2 \\ &= \sum_{i=1}^n h_{i,i}^2 + 2\sum_{i>j} h_{ij}^2.\end{aligned}$$

Thus, we can write the density with respect to the above measure as

$$f(H) = \prod_{i=1}^n \frac{1}{\sqrt{4\pi}} e^{-\frac{h_{ii}^2}{4}} dh_{ii} \prod_{i>j} \frac{1}{\sqrt{2\pi}} e^{-\frac{h_{ij}^2}{2}} dh_{ij}.$$

Hence, the normalising constant is $Z_n = (4\pi)^{\frac{n}{2}}(2\pi)^{\frac{1}{2}\binom{n}{2}}$. In general, Gaussian ensembles can be written as

$$f_\beta(H) = \frac{1}{Z_{n,\beta}} \exp[-\frac{\beta}{4}\mathrm{tr}H^2],$$

where $\beta = 1, 2, 4$ corresponding to Gaussian Orthogonal Ensemble i.e., GOE (when we consider random matrix with real valued random variables), Gaussian Unitary Ensemble i.e., GUE (when we consider random matrix with complex valued random variables) or Gaussian Symplectic Ensemble i.e., GSE (when considered over quaternions). A GOE is invariant under an orthogonal conjugation. Given a GOE, H_n and an $(n \times n)$ orthogonal

matrix P, then $P^T HP$ has the same density as H_n. Mehta [33] gives the result that the joint density of the eigenvalues of a GOE/GUE/GSE

$$f_\beta(l_1, l_2, \ldots, l_n) = C_{N,\beta} \prod_{i<j} |l_i - l_j|^\beta e^{-\frac{\beta}{2}\sum_{i=1}^n l_i^2} \tag{3.12}$$

where $C_{n,\beta}$ is a known normalising constant and $l_1 > l_2 > \ldots > l_n$. It is interesting to note that the joint eigenvalue distributions corresponding to the three separate cases can be expressed in a unified notation parameterised by β. Tracy and Widom [46, 47] derived the asymptotic distribution of the largest eigenvalue for the above three canonical Gaussian ensembles. They provide limit laws as the matrix dimension goes to infinity. If

$$F_{n,\beta}(t) = P_{n,\beta}(l_1 < t), \quad \beta = 1, 2, 4$$

denotes the distribution function of the largest eigenvalue, then the existence and the closed form expressions for the three get from [46, 47, 48] that

$$F_\beta(x) = \lim_{n\to\infty} F_{n,\beta}(2\sigma\sqrt{n} + \frac{\sigma x}{n^{\frac{1}{6}}}) \tag{3.13}$$

exist and is given by

$$F_2(x) = \exp\left(-\int_x^\infty (y-x)q^2(y)dy\right)$$
$$F_1(x) = (F_2(x))^{\frac{1}{2}} \exp\left(-\frac{1}{2}\int_x^\infty q(y)dy\right)$$
$$F_4(x) = (F_2(x))^{\frac{1}{2}} \cosh\left(-\frac{1}{2}\int_x^\infty q(y)dy\right)$$

where q is the unique solution to the *Painlevè II equation* $q'' = xq + 2q^3$ satisfying the boundary condition $q(x) \sim \mathrm{Ai}(x)$ as $x \to \infty$, and where $\mathrm{Ai}(x)$ denotes the Airy function.

Expressions for the orthogonal and symplectic ensembles for $n \to \infty$ are presented in [48] as well as the tail behaviour when $x \to +\infty$. The results corresponding to the case of $x \to -\infty$ can be found in [50]. Johnstone [24] proved a universality result under the null hypothesis that $\Sigma = I_p$ and showed that under appropriate shifting and scaling, the asymptotic distribution of the largest eigenvalue of Wishart matrices converges to the Tracy-Widom limit.

3.2.1 Gaussian Orthogonal Ensemble

In this subsection, we derive the exact cumulative distribution function of the largest eigenvalue of a Gaussian orthogonal ensemble. It will be evident from the proof of the next theorem that Gaussian unitary and symplectic ensembles may be developed in exactly the same manner.

Theorem 3.2. *The cdf of the largest eigenvalue for a $N \times N$ GOE is $P(l_1 \leq x) = C_N \Omega$ where $\Omega = Pf(A)$. $Pf(A)$ denotes the Pfaffian of a matrix A and $A = (a_{ij})$ is a skew-symmetric matrix such that and $a_{ij} = I^1_{ij} - I^2_{ij}$, where*

$$I^1_{ij} = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{\Gamma(c+1)}{2^n \Gamma(a+n+1)} \frac{x^{2k}}{\Gamma(c+k+1)} - \sum_{n=0}^{\infty} \frac{\Gamma(c)}{2^n \Gamma(a+n+1)} - \frac{1}{2}\Gamma(a)\Gamma(b)[x^{2b} e^{-\frac{x^2}{2}} \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(b+n+1)} - 1].$$

and

$$I^2_{ij} = \frac{1}{4}\Gamma(a)\Gamma(b) \left[x^{2a} e^{-\frac{x^2}{2}} \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(a+n+1)} - 1 \right] \times \left[x^{2b} e^{-\frac{x^2}{2}} \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(b+n+1)} - 1 \right].$$

where $a = i/2, b = j/2, \alpha = i + j + 2n$ and $c = \alpha/2$.

Proof. Let $l_1 = l_{max}$ denote the largest eigenvalue. Hence, the distribution function of the largest eigenvalue is given by

$$\begin{aligned} F(x) &= P(l_1 \leq x) \\ &= P(l_1 \leq x, l_2 \leq x, \ldots, l_N \leq x) \\ &= C_N \Omega. \end{aligned}$$

Leaving the normalising constant, the integral that we need to compute is given by

$$\Omega = \int\limits_{[-\infty,x]^N} \prod_{i<j} |l_i - l_j|^\beta e^{-\frac{\beta}{2}\sum_{i=1}^N l_i^2} dl_1 dl_2 \ldots dl_N.$$

Setting $\beta = 1$ for GOE we get,

$$\Omega = \int\limits_{[-\infty,x]^N} \prod_{i<j} |l_i - l_j| e^{-\frac{1}{2}\sum\limits_{i=1}^N l_i^2} dl_1 dl_2 \ldots dl_N.$$

Note that the joint distribution of the eigenvalues is the same as the joint distribution of the order statistics of the eigenvalues, except the normalising constant. Further $\prod_{i<j} |l_i - l_j| = \det(l_j^{i-1})$ and since we consider that eigenvalues are ordered from the largest to the smallest, we denote l_1 to be the largest eigenvalue and l_N to be the smallest eigenvalue.

Let the Stieltjes' measure be defined as $d\mu(l) = \exp[-\frac{1}{2}l^2]dl$. Let $g(l) = \exp[-\frac{1}{2}l^2]$. It is easily seen that $g(l)$ is increasing when $l < 0$ and $g(l)$ is decreasing when $l > 0$. Let us suppose that $x < 0$ so that we find the distribution function of l_1 when $l_1 < 0$. The case for $x > 0$ can be handled analogously. Thus,

$$\begin{aligned} \Omega &= \int\limits_{[-\infty,x]^N} \prod_{i<j} |l_i - l_j| d\mu(l_1) d\mu(l_2) \ldots d\mu(l_N) \\ &= \int\limits_{[-\infty,x]^N} \det(\phi_i(l_j)) d\mu(l_1) d\mu(l_2) \ldots d\mu(l_N). \end{aligned}$$

Now, using De Bruijn's theorem, we know that $\Omega = Pf(A)$ where A is a skew-symmetric matrix(of even order for the time being). Let a_{ij} denote the $(i,j)^{th}$ entry of the matrix. Hence,

$$\Omega = Pf\left[\int_{-\infty}^{x}\int_{-\infty}^{x} \text{sgn}(y-t)\phi_i(t)\phi_j(y)d\mu(t)d\mu(y)\right].$$

Writing the $(i,j)^{th}$ entry of the matrix A, we get,

$$\begin{aligned}
a_{ij} &= \int_{-\infty}^{x}\int_{-\infty}^{x} \text{sgn}(y-t)\phi_i(t)\phi_j(y)d\mu(t)d\mu(y) \\
&= \int_{-\infty}^{x}\int_{-\infty}^{x} \text{sgn}(y-t)t^{i-1}y^{j-1}d\mu(t)d\mu(y) \\
&= \int_{-\infty}^{x}\int_{-\infty}^{y} y^{j-1}t^{i-1}d\mu(t)d\mu(y) - \int_{-\infty}^{x}\int_{y}^{x} y^{j-1}t^{i-1}d\mu(t)d\mu(y) \\
&= 2\int_{-\infty}^{x}\int_{-\infty}^{y} t^{i-1}y^{j-1}d\mu(t)d\mu(y) - \int_{-\infty}^{x}\int_{-\infty}^{x} t^{i-1}y^{j-1}d\mu(t)d\mu(y) \\
&= I^1_{i,j} - I^2_{i,j}. \qquad (3.14)
\end{aligned}$$

where,

$$I^1_{i,j} = 2\int_{-\infty}^{x}\int_{-\infty}^{y} t^{i-1}y^{j-1}d\mu(t)d\mu(y). \qquad (3.15)$$

and

$$I^2_{i,j} = \int_{-\infty}^{x}\int_{-\infty}^{x} t^{i-1}y^{j-1}d\mu(t)d\mu(y). \qquad (3.16)$$

Let us look at (3.16) first

$$\begin{aligned} I_{i,j}^2 &= \int_{-\infty}^{x}\int_{-\infty}^{x} t^{i-1}y^{j-1}d\mu(t)d\mu(y) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{x} t^{i-1}y^{j-1}e^{-\frac{t^2}{2}}e^{-\frac{y^2}{2}}dtdy \\ &= \int_{-\infty}^{x} t^{i-1}e^{-\frac{t^2}{2}}dt \int_{-\infty}^{x} y^{j-1}e^{-\frac{y^2}{2}}dy. \end{aligned}$$

Consider

$$I_{ij}^{2'} = \int_{-\infty}^{x} t^{i-1}e^{-\frac{t^2}{2}}dt.$$

Substituting $\frac{t^2}{2} = v$ we rewrite the above as

$$\begin{aligned} &= 2^{\frac{i}{2}-1}\int_{\infty}^{\frac{x^2}{2}} v^{\frac{i}{2}-1}e^{-v}dv \\ &= -2^{\frac{i}{2}-1}\int_{\frac{x^2}{2}}^{\infty} v^{\frac{i}{2}-1}e^{-v}dv \\ &= -2^{\frac{i}{2}-1}\Gamma(a,\frac{x^2}{2}). \end{aligned}$$

where $a = \frac{i}{2}$ and the last term refers to an upper incomplete Gamma function, see [56]. We know from [56] that $\Gamma(a,\frac{x^2}{2}) = \Gamma(a) - \gamma(a,\frac{x^2}{2})$ where $\gamma(a,\frac{x^2}{2})$ is the lower incomplete Gamma function and

$$\gamma(a,\frac{x^2}{2}) = a^{-1}(\frac{x^2}{2})^a e^{-\frac{x^2}{2}}F_1(1,a+1,\frac{x^2}{2})$$

where F_1 is the confluent Hypergeometric series given by

$$F(a,c,z) = \sum_{n=0}^{\infty}\frac{(a)_n}{(c)_n}\frac{z^n}{n!}$$

for $c \neq 0, -1, -2, \ldots$

$$= \frac{\Gamma(c)}{\Gamma(a)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)}{\Gamma(c+n)} \frac{z^n}{n!}.$$

Now,

$$\begin{aligned} F(1, a+1, \frac{x^2}{2}) &= \sum_{n=0}^{\infty} \frac{(1)_n}{(a+1)_n} \frac{x^{2n}}{2^n n!} \\ &= \frac{\Gamma(a+1)}{\Gamma(1)} \sum_{n=0}^{\infty} \frac{\Gamma(n+1)}{\Gamma(a+n+1)} \frac{(x^2)^n}{2^n n!} \\ &= \Gamma(a+1) \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(a+n+1)}. \end{aligned}$$

Thus we get,

$$\gamma(a, \frac{x^2}{2}) = \frac{x^{2a}}{a \cdot 2^a} e^{-\frac{x^2}{2}} \Gamma(a+1) \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(a+n+1)}. \tag{3.17}$$

The infinite series given in (3.17) converges for all x since $\Gamma(a+n+1)$ is non-negative. We therefore get

$$\begin{aligned} I_{ij}^{2'} &= \int_{-\infty}^{x} t^{i-1} e^{-\frac{t^2}{2}} dt \\ &= -2^{\frac{i}{2}-1} \Gamma(a, \frac{x^2}{2}) \\ &= -2^{\frac{i}{2}-1} [\Gamma(a) - \gamma(a, \frac{x^2}{2}] \\ &= 2^{\frac{i}{2}-1} [\gamma(a, \frac{x^2}{2}) - \Gamma(a)] \\ &= 2^{\frac{i}{2}-a-1} \Gamma(a) \left[x^{2a} e^{-\frac{x^2}{2}} \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(a+n+1)} - 1 \right] \\ &= \frac{1}{2} \Gamma(a) \left[x^{2a} e^{-\frac{x^2}{2}} \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(a+n+1)} - 1 \right]. \end{aligned} \tag{3.18}$$

Similarly, letting $b = \frac{j}{2}$ we can write

$$\begin{aligned} I_{ij}^{2''} &= \int_{-\infty}^{x} y^{j-1} e^{-\frac{y^2}{2}} dy \\ &= 2^{\frac{j}{2}-b-1}\Gamma(b)\left[x^{2b}e^{-\frac{x^2}{2}}\sum_{n=0}^{\infty}\frac{x^{2n}}{2^n\Gamma(b+n+1)} - 1\right] \\ &= \frac{1}{2}\Gamma(b)\left[x^{2b}e^{-\frac{x^2}{2}}\sum_{n=0}^{\infty}\frac{x^{2n}}{2^n\Gamma(b+n+1)} - 1\right]. \end{aligned} \tag{3.19}$$

We thus get $I_{i,j}^2 = (3.18) \times (3.19)$. Now, let's look at $I_{i,j}^1$. From (3.15) we see,

$$\begin{aligned} I_{i,j}^1 &= 2\int_{-\infty}^{x}\int_{-\infty}^{y} t^{i-1}y^{j-1}d\mu(t)d\mu(y) \\ &= 2\int_{-\infty}^{x}\int_{-\infty}^{y} t^{i-1}y^{j-1}e^{-\frac{t^2}{2}}e^{-\frac{y^2}{2}}dtdy \end{aligned}$$

$$\begin{aligned} I_{i,j}^1 &= 2\int_{-\infty}^{x} y^{j-1}e^{-\frac{y^2}{2}}\left[\int_{-\infty}^{y} t^{i-1}e^{-\frac{t^2}{2}}dt\right]dy \\ &= \Gamma(a)\int_{-\infty}^{x} y^{j-1}e^{-\frac{y^2}{2}}\left[y^{2a}e^{-\frac{y^2}{2}}\sum_{n=0}^{\infty}\frac{y^{2n}}{2^n\Gamma(a+n+1)} - 1\right]dy \\ &= \Gamma(a)\int_{-\infty}^{x} y^{2a+j-1}e^{-y^2}\sum_{n=0}^{\infty}\frac{y^{2n}}{2^n\Gamma(a+n+1)}dy - \Gamma(a)\int_{-\infty}^{x} y^{j-1}e^{-\frac{y^2}{2}}dy \\ &= \Gamma(a)[I_{ij}^{1'} - I_{ij}^{1''}] \end{aligned}$$

where

$$I_{ij}^{1'} = \int_{-\infty}^{x}\sum_{n=0}^{\infty}\frac{y^{2a+j+2n-1}e^{-y^2}}{2^n\Gamma(a+n+1)}dy.$$

Let $\alpha = 2a + j + 2n$, since the infinite sum is convergent, we can take it out of the integral to get

$$
\begin{aligned}
&= \sum_{n=0}^{\infty} \frac{1}{2^n \Gamma(a+n+1)} \int_{-\infty}^{x} y^{\alpha-1} e^{-y^2} dy \\
&= -\sum_{n=0}^{\infty} \frac{1}{2^n \Gamma(a+n+1)} \int_{x^2}^{\infty} v^{\frac{\alpha}{2}-1} e^{-v} dv \\
&= -\sum_{n=0}^{\infty} \frac{1}{2^n \Gamma(a+n+1)} \Gamma(c, x^2).
\end{aligned}
$$

where $c = \frac{\alpha}{2}$

$$
= -\sum_{n=0}^{\infty} \frac{\Gamma(c)}{2^n \Gamma(a+n+1)} + \sum_{n=0}^{\infty} \frac{\gamma(c, x^2)}{2^n \Gamma(a+n+1)} \tag{3.20}
$$

Now, remember that

$$
\begin{aligned}
\gamma(c, x^2) &= \frac{x^{2c}}{c} e^{-x^2} F_1(1, c+1, x^2) \\
F_1(1, c+1, x^2) &= \Gamma(c+1) \sum_{n=0}^{\infty} \frac{x^{2k}}{\Gamma(c+k+1)}.
\end{aligned}
$$

Substituting the last two expressions in (3.20), we get

$$
\sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{\Gamma(c+1)}{2^n \Gamma(a+n+1)} \frac{x^{2k}}{\Gamma(c+k+1)} - \sum_{n=0}^{\infty} \frac{\Gamma(c)}{2^n \Gamma(a+n+1)}. \tag{3.21}
$$

Therefore we have

$$
\begin{aligned}
I_{ij}^{1''} &= \Gamma(a) \int_{-\infty}^{x} y^{j-1} e^{-\frac{y^2}{2}} \\
&= \frac{1}{2} \Gamma(a) \Gamma(b) [x^{2b} e^{-\frac{x^2}{2}} \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n \Gamma(b+n+1)} - 1].
\end{aligned} \tag{3.22}
$$

Thus, setting $I_{ij}^1 =$ (3.21)-(3.22) we get the required integral. Consequently, we have closed form analytical expressions to get the $(i,j)^{th}$ entry of the matrix A as $a_{ij} = I_{ij}^1 - I_{ij}^2$.

The required integral is given by $\Omega = Pf(A)$ and consequently the cumulative distribution function of the largest eigenvalue for the GOE is given by $F(x) = C_{N,1}\Omega$. □

3.2.2 Extensions to GUE/GSE

The above approach can be easily extended to the cases of $\beta = 2, 4$ as well. For $\beta = 2$, that is for Gaussian Unitary Ensemble , the main integral to be evaluated is

$$\Omega = \int_{[-\infty,x]^N} \prod_{i<j}(l_i - l_j)^2 e^{-\sum_{i=1}^N l_i^2} dl_1 dl_2 \dots dl_N.$$

The only difference in the entire analysis from the GOE case corresponds to the treatment of the Vandermonde term. Using (3.9) with $\phi_i(x) = x^i, \psi_j(x) = x^j$, we can write $\prod_{i<j}(l_i - l_j)^2 = \det(\phi_i(x)\psi_j(x))$. As a consequence of Theorem 3.1[3.9], the above multiple integral evaluation reduces to the following evaluation of a determinant whose $(i,j)^{th}$ entries are single integrals as shown below with,

$$a_{ij} = \int_{-\infty}^{x} \phi_i(x)\psi_j(x)d\mu(x)$$

$\Omega = \det(A)$ where $A = (a_{ij})$. The principle analysis would in fact be simpler since we would be dealing with single integrals and hence with single incomplete gamma functions.

For the case of $\beta = 4$, the Guassian Symplectic Ensemble, the integral to be evaluated is

$$\Omega = \int_{[-\infty,x]^N} \prod_{i<j}(l_i - l_j)^4 e^{-2\sum_{i=1}^N l_i^2} dl_1 dl_2 \dots dl_N.$$

It is to be noted that the primary difference in the above integral once again happens to be the Vandermonde term, which in this case is $\prod_{i<j}(l_i - l_j)^4$ which can be dealt with as in Dieng and Tracy [12] by appealing to

$$\prod_{0\le j<k\le N} (x_j - x_k)^4 = \det(x_k^j \quad jx_k^{j-1})_{j=0,\dots 2N-1, k=1,\dots,N}, \tag{3.23}$$

and applying Theorem 3.1[3.11]. The exact evaluations are somewhat more direct in the $\beta = 4$ case since one does not need to deal with the sgn$(\cdot)$ term in (3.14) as in the $\beta = 1$ setting.

3.3 Results for Jacobi Ensemble

The Double Wishart setting is also referred to in literature as the Jacobi ensemble owing to its representation in terms of the Jacobi polynomials. Again, we only present the $\beta = 1$ case, i.e., real valued matrices. The definition of the *greatest root statistic* was given in section (3.1.2). An approximate expression in terms of a *hypergeometric function with a matrix argument* was also presented. Koev et al. in [29] present efficient algorithms and implementation using MATLAB to evaluate a hypergeometric function with a matrix argument. Johnstone [25] reports that current MATLAB evaluations to compute the distribution function of the greatest root statistic take about 1 second for $n_1, n_2, p \leq 17$. Recently, Johnstone in [25] presented an asymptotic result of the greatest root statistic. The universality behaviour of the largest eigenvalue of this class of random matrices was established. It turns out that under some growth conditions on the sample sizes, n_1, n_2 and the number of variables, p, the logit transform of the greatest root statistic also follows the Tracy-Widom law. More precisely, assume p is even and that $p, n_1(p)$ and $n_2(p)$ tend to infinity together in such a way that

$$\lim_{p\to\infty} \frac{\min(p, n_2)}{n_1 + n_2} > 0, \qquad \lim_{p\to\infty} \frac{p}{n_1} < 1.$$

Then the following is true.

Theorem 3.3 (Johnstone, 2008). *Let $l_1(p)$ denote the greatest root statistic. Assume that $n_1, n_2 \to \infty$ as $p \to \infty$ through even values of p as given above. For each $s_0 \in \mathbb{R}$, there exists*

$C > 0$ such that for all $s \geq s_0$,

$$|P\{W_p \leq \mu_p + \sigma_p s\} - F_1(s)| \leq Cp^{-\frac{2}{3}} e^{-\frac{s}{2}},$$

where $F_1(s)$ is the CDF of the Tracy-Widom distribution, $W_p = \log \frac{l_1(p)}{1-l_1(p)}$, μ_p and σ_p are centring and scaling constants given by

$$\mu_p = 2\log\tan(\frac{\phi+\gamma}{2}) \quad \textit{and} \quad \sigma_p^3 = \frac{16}{(n_1+n_2-1)^2}\frac{1}{\sin^2(\phi+\gamma)\sin\phi\sin\gamma},$$

where the angle parameters ϕ and γ depend on n_1, n_2 and p.

We are however interested in deriving the exact cumulative distribution function of the greatest root statistic which effectively reduces to evaluate the following integral.

$$\Omega = \int_{[0,x]^p} \prod_{i=1}^{p} l_i^{b_1}(1-l_i)^{b_2} \prod_{i<j} |l_i - l_j| dl_1 dl_2 \ldots dl_p. \tag{3.24}$$

Theorem 3.4. *The integral in (3.24) is $\Omega = Pf(A)$, where $Pf(A)$ denotes the Pfaffian of a matrix A, and $A = (a_{ij})$ is a skew-symmetric matrix whose $(i,j)^{th}$ entry is given by $a_{ij} = I_{ij}^1 - I_{ij}^2$ where*

$$I_{ij}^1 = 2\sum_{n=0}^{b_2}\sum_{k=0}^{b_2} \frac{(-b_2)_n(-b_2)_k}{(i+b_1+n)(\alpha+n+k)} \frac{x^{\alpha+n+k}}{(n!)(k!)}$$

$$I_{ij}^2 = \left[x^{i+b_1}\Gamma(b_2+1)\sum_{n=0}^{b_2}(-1)^n \frac{1}{(i+b_1+n)\Gamma(b_2+1-n)}\frac{x^n}{n!}\right] \left[x^{j+b_1}\Gamma(b_2+1)\sum_{n=0}^{b_2}(-1)^n \frac{1}{(j+b_1+n)\Gamma(b_2+1-n)}\frac{x^n}{n!}\right]$$

where $(-b_2)_n = (-1)^n(b_2-n+1)_n$, $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}$ and $\alpha = i+j+2b_1$.

Proof. Suppose $l_1 \geq l_2 \geq \ldots l_p$ are the eigenvalues of the multivariate beta distribution. Then, owing to the positive definiteness of the Wishart matrices used in the construciton of

the multivariate beta matrix, all the eigenvalues of the multivariate beta matrix lie between 0 and 1. Now,

$$\begin{aligned} P(l_1 \leq x) &= P(l_1 \leq x, l_2 \leq x, \ldots, l_p \leq x) \\ &= \int\limits_{[0,x]^p} f(\mathbf{l}) d\mathbf{l}. \end{aligned}$$

Denoting the joint distribution by J_1

$$P(J_1 \leq x) = \int\limits_{[0,x]^p} f(\mathbf{l}) d\mathbf{l}.$$

As a reminder, $a_1 = n_1/2$ and $b_1 = n_2/2$. Let $b_1 = a_1 - \frac{p-1}{2} - 1$ and $b_2 = a_2 - \frac{p-1}{2} - 1$. As such, we have $b_1 = \frac{n_1-p+1}{2} - 1$ and $b_2 = \frac{n_2-p+1}{2} - 1$. The normalising constant in the joint density function is completely specified for known values of a_1 and a_2. We need an analytical expression for the integral given above. We use Theorem 3.1 (ii). Note that the joint distribution of the eigenvalues is the same as the joint distribution of the order statistics of the eigenvalues. Further, $\prod_{i<j} |l_i - l_j| = \det(l_j^{i-1})$ For the time being, let us assume that the matrix is of even order. The odd order case would require just a slight modification. The integral that we need to evaluate is given by

$$\begin{aligned} \Omega &= \int\limits_{[0,x]^p} \prod_{i=1}^{p} l_i^{a_1 - \frac{p-1}{2} - 1} (1 - l_i)^{a_2 - \frac{p-1}{2} - 1} \prod_{i<j} |l_i - l_j| dl_1 dl_2 \ldots dl_p \\ &= \int\limits_{[0,x]^p} \prod_{i=1}^{p} l_i^{b_1} (1 - l_i)^{b_2} \prod_{i<j} |l_i - l_j| dl_1 dl_2 \ldots dl_p. \end{aligned} \tag{3.25}$$

Define the following Stieltjes' measure as $\mu(l) = l^{b_1}(1-l)^{b_2} dl$. Let $g(l) = l^{b_1}(1-l)^{b_2}$. We need the above measure to be a valid Stieltjes' measure so that we can use De Bruijn's Theorem and as such we need conditions for which $g(l)$ would be a monotone function. Thus,

$$g(l) = l^{b_1}(1 - l)^{b_2}.$$

Differentiating with respect to l we get

$$\begin{aligned} g'(l) &= b_1 l^{b_1-1}(1-l)^{b2} - b_2 l^{b_1}(1-l)^{b_2-1} \\ &= l^{b_1-1}(1-l)^{b_2-1}[b_1(1-l) - b_2 l]. \end{aligned}$$

We know that b_1, b_2 are positive. Let us further assume that $b_1, b_2 > 1$, which is equivalent to saying that $n_1 > p+2$ and $n_2 > p+2$. We know that $l \in [0,1]$. Therefore, $0 < l^{b_1-1}(1-l)^{b_2-1} < 1 \quad \forall l$. For $g(l)$ to be a monotone increasing function, we need $g'(l) > 0$. This implies $b_1(1-l) - b_2 l > 0$ and so that $l < \frac{b_1}{b_1+b_2}$. Therefore, if $x < \frac{b_1}{b_1+b_2}$, then we can use the Stieltjes' measure, for $g(l)$ would be a monotone increasing function in that case. Note also that $g(\frac{b_1}{b_1+b_2}) = 0$ and the function is symmetric around $\frac{b_1}{b_1+b_2}$ so for $x > \frac{b_1}{b_1+b_2}$, the function is monotone decreasing. As such, for this case the integral computation would be symmetric with a negative sign. Let us consider the case of $x < \frac{b_1}{b_1+b_2}$. Let $d\mu(l_j) = l_j^{b_1}(1-l_j)^{b_2} dl$ and define

$$\begin{aligned} \Omega &= \int_{[0,x]^p} \prod_{i=1}^{p} l_i^{b_1}(1-l_i)^{b_2} \prod_{i<j} |l_i - l_j| dl_1 dl_2 \dots dl_p \\ &= \int_{[0,x]^p} \prod_{i<j} (l_i - l_j) d\mu(l_1) d\mu(l_2) \dots d\mu(l_p) \\ &= \int_{[0,x]^p} \det(\phi_i(l_j)) d\mu(l_1) d\mu(l_2) \dots d\mu(l_p). \end{aligned} \tag{3.26}$$

As written earlier, $\phi_i(l_j) = l_j^{i-1}$ and the determinant thus computed is the classical Vandermonde determinant. Now, using De Bruijn's Theorem, we know that $\Omega = Pf(A)$ where A is a skew-symmetric matrix (of even order for the time being), let a_{ij} denote the $(i,j)^{th}$ entry

of the matrix. From Theorem 3.1 (ii) we get,

$$\begin{aligned} a_{ij} &= \int_0^x \int_0^x \phi_i(t)\phi_j(y)\text{sgn}(y-t)d\mu(t)d\mu(y) \\ &= \int_0^x \int_0^x t^{i-1}y^{j-1}\text{sgn}(y-t)d\mu(t)d\mu(y) \\ &= \int_0^x \int_0^y t^{i-1}y^{j-1}d\mu(t)d\mu(y) - \int_0^x \int_y^x t^{i-1}y^{j-1}d\mu(t)d\mu(y) \end{aligned}$$

$$\begin{aligned} a_{ij} &= 2\int_0^x \int_0^y t^{i-1}y^{j-1}d\mu(t)d\mu(y) \\ &\quad - \left[\int_0^x \int_0^y t^{i-1}y^{j-1}d\mu(t)d\mu(y) + \int_0^x \int_y^x t^{i-1}y^{j-1}d\mu(t)d\mu(y)\right] \\ &= I_{ij}^1 - I_{ij}^2 \end{aligned}$$

where

$$I_{ij}^1 = 2\int_0^x \int_0^y t^{i-1}y^{j-1}d\mu(t)d\mu(y) \tag{3.27}$$

and

$$\begin{aligned} I_{ij}^2 &= \int_0^x \int_0^y t^{i-1}y^{j-1}d\mu(t)d\mu(y) + \int_0^x \int_y^x t^{i-1}y^{j-1}d\mu(t)d\mu(y) \\ &= \int_0^x \int_0^x t^{i-1}y^{j-1}d\mu(t)d\mu(y). \end{aligned} \tag{3.28}$$

Let us first look at (3.28), it follows that

$$\begin{aligned} I_{ij}^2 &= \int_0^x \int_0^x t^{i-1} y^{j-1} d\mu(t) d\mu(y) \\ &= \int_0^x \int_0^x t^{i-1} y^{j-1} t^{b_1} y^{b_1} (1-t)^{b_2} (1-y)^{b_2} dt dy \\ &= \left[\int_0^x t^{i+b_1-1} (1-t)^{b_2} dt \right] \left[\int_0^x y^{j+b_1-1} (1-y)^{b_2} dy \right] \\ &= B(i+b_1, b_2+1, x) B(j+b_1, b_2+1, x). \end{aligned} \tag{3.29}$$

where $B(\cdot, \cdot, x)$ is the incomplete beta function. From [56], we see that incomplete beta functions can be represented as a Gaussian hypergeometric series. The standard Gaussian Hypergeometric series is given by

$$F(a, b, c, z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!},$$

where $(a)_n$ is called the Pochhammer symbol given by $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}$. We also note from [56] that if in the Gaussian Hypergeometric series, either a or b is a negative integer, then the series becomes a polynomial, that is,

$$F(-m, b, c, z) = \sum_{n=0}^{m} \frac{(-m)_n (b)_n}{(c)_n} \frac{z^n}{n!}.$$

Given an incomplete Beta function as

$$B(a, b, z) = \int_0^z t^{a-1} (1-t)^{b-1} dt$$

with $0 < z < 1$ we can write it as

$$B(a, b, z) = \frac{z^a}{a} F(a, 1-b, a+1, z).$$

where $F(a, 1-b, a+1, z)$ is the standard Gaussian hypergeometric series. In our case, we see that,

$$\begin{aligned} B(i+b_1, b_2+1, x) &= \frac{x^{i+b_1}}{i+b_1} F(i+b_1, -b_2, i+b_1+1, x) \\ &= \frac{x^{i+b_1}}{i+b_1} \sum_{n=0}^{b_2} \frac{(i+b_1)_n(-b_2)_n}{(i+b_1+1)_n} \frac{x^n}{n!}. \end{aligned} \tag{3.30}$$

We further have, $(-b_2)_n = (-1)^n(b_2-n+1)_n$ so that, we can rewrite (3.30) as

$$B(i+b_1, b_2+1, x) = \frac{x^{i+b_1}}{i+b_1} \sum_{n=0}^{b_2} (-1)^n \frac{(i+b_1)_n(b_2-n+1)_n}{(i+b_1+n)_n} \frac{x^n}{n!} \tag{3.31}$$

$$\begin{aligned} &= \frac{x^{i+b_1}}{i+b_1} \sum_{n=0}^{b_2} (-1)^n \frac{i+b_1}{i+b_1+n} \frac{\Gamma(b_2+1)}{\Gamma(b_2+1-n)} \frac{x^n}{n!} \\ &= x^{i+b_1}\Gamma(b_2+1) \\ &\qquad \sum_{n=0}^{b_2} (-1)^n \frac{1}{(i+b_1+n)\Gamma(b_2+1-n)} \frac{x^n}{n!}. \end{aligned} \tag{3.32}$$

Using (3.31), we can write (3.29) as

$$\begin{aligned} I_{ij}^2 = &\left[x^{i+b_1}\Gamma(b_2+1) \sum_{n=0}^{b_2} (-1)^n \frac{1}{(i+b_1+n)\Gamma(b_2+1-n)} \frac{x^n}{n!} \right] \\ &\left[x^{j+b_1}\Gamma(b_2+1) \sum_{n=0}^{b_2} (-1)^n \frac{1}{(j+b_1+n)\Gamma(b_2+1-n)} \frac{x^n}{n!} \right]. \end{aligned} \tag{3.33}$$

Now, upon examination of (3.27), which is denoted by I_{ij}^1 we see

$$\begin{aligned} I_{ij}^1 &= 2 \int_0^x \int_0^y t^{i-1} y^{j-1} d\mu(t) d\mu(y) \\ &= 2 \int_0^x y^{j-1} \left[\int_0^y t^{i-1} d\mu(t) \right] d\mu(y). \end{aligned}$$

Now the inner integral is an incomplete beta function that is given by

$$\begin{aligned} \int_0^y t^{i-1} d\mu(t) &= \int_0^y t^{i+b_1-1}(1-t)^{b_2} dt \\ &= \frac{y^{i+b_1}}{i+b_1} F(i+b_1, -b_2, i+b_1+1, y). \end{aligned}$$

Thus, we can write I_{ij}^1 as

$$I_{ij}^1 = \frac{2}{i+b_1}\int_0^x y^{i+j+2b_1-1}(1-y)^{b_2}F(i+b_1,-b_2,i+b_1+1,y)dy.$$

Let $\alpha = i+j+2b_1$, then

$$I_{ij}^1 = \frac{2}{i+b_1}\int_0^x y^{\alpha-1}(1-y)^{b_2}F(i+b_1,-b_2,i+b_1+1,y)dy.$$

Since $i,j,b_1,b_2 \geq 1$ we have,

$$= \frac{2}{i+b_1}\int_0^x y^{\alpha-1}(1-y)^{1+b_2-1}\left(\sum_{n=0}^{b_2}\frac{(i+b_1)_n(-b_2)_n}{(i+b_1+1)_n}\frac{y^n}{n!}\right)dy.$$

The summation is a polynomial and hence taking it outside of the integral

$$\begin{aligned}
&= \frac{2}{i+b_1}\sum_{n=0}^{b_2}\frac{(i+b_1)(-b_2)_n}{(i+b_1+n)(n!)}\int_0^x y^{\alpha+n-1}(1-y)^{1+b_2-1}dy \\
&= 2\sum_{n=0}^{b_2}\frac{(-b_2)_n}{(i+b_1+n)(n!)}B(\alpha+n,1+b_2,x) \\
&= 2\sum_{n=0}^{b_2}\frac{(-b_2)_n}{(i+b_1+n)(n!)}\frac{x^{\alpha+n}}{(\alpha+n)}F(\alpha+n,-b_2,\alpha+n+1,x) \\
&= 2\sum_{n=0}^{b_2}\frac{(-b_2)_n x^{\alpha+n}}{(i+b_1+n)(\alpha+n)(n!)}\sum_{k=0}^{b_2}\frac{(\alpha+n)_k(-b_2)_k}{(\alpha+n+1)_k}\frac{x^k}{k!} \\
&= 2\sum_{n=0}^{b_2}\frac{(-b_2)_n x^{\alpha+n}}{(i+b_1+n)(\alpha+n)(n!)}\sum_{k=0}^{b_2}\frac{(\alpha+n)(-b_2)_k}{(\alpha+n+k)}\frac{x^k}{k!} \\
&= 2\sum_{n=0}^{b_2}\sum_{k=0}^{b_2}\frac{(-b_2)_n(-b_2)_k}{(i+b_1+n)(\alpha+n+k)}\frac{x^{\alpha+n+k}}{(n!)(k!)}. \qquad (3.34)
\end{aligned}$$

Thus, using (3.33) and (3.34) we can write the $(i,j)^{th}$ entry of the matrix A as a_{ij} = (3.34) − (3.33). □

Proposition 3.1. *For the real Jacobi case as above, the cdf of the largest eigenvalue is obtained as the Pfaffian of a skew-symmetric matrix whose $(i,j)^{th}$ entry is derived above.*

Let $A = (a_{ij})$ denote the required skew matrix. $P(\lambda_1 \leq x) = C_{p,1}Pf(A)$, where $C_{p,1}$ is a normalizing constant. We show that $0 \leq Pf(A) \leq 1$.

Proof. Note that, we are considering only the matrix of even order for the time being. For a matrix of even order, its determinant can be written as the square of a polynomial in the matrix entries. This polynomial is called the Pfaffian of the matrix. Thus, $Pf(A) = \sqrt{\det(A)}$. We first show that a_{ij} for each (i, j) is lies between -1 and 1 and strictly nonzero for some i and j in the range of integration. From above, we see that

$$a_{ij} = I^1_{ij} - I^2_{ij}$$

where I^1_{ij} and I^2_{ij} are given as

$$I^1_{ij} = 2\int_0^x \int_0^y t^{i-1}y^{j-1}d\mu(t)d\mu(y)$$
$$I^2_{ij} = \int_0^x \int_0^x t^{i-1}y^{j-1}d\mu(t)d\mu(y).$$

Obviously $y \leq x$. Thus, $I^1_{ij} \leq 2I^2_{ij}$. Hence, $(I^1_{ij} - I^2_{ij}) \leq I^2_{ij}$. Note that

$$I^2_{ij} = B(i + b_1, b_2 + 1, x)B(j + b_1, b_2 + 1, x),$$

which by definition they lies between 0 and 1, therefore, $I^2_{ij} < 1$ for $x < 1$. Consequently, $a_{ij} < 1$ for every value of i and j.

We thus see that each entry of the skew symmetric matrix lies between -1 and 1. It is known that the determinant of a matrix can be interpreted as the area of the parallelogram whose vertices are given by the columns of the matrix. Consider a skew matrix $B = (b_{ij})$ of

order $(2n \times 2n)$ where $n \in I^+$ such that $b_{ji} = -b_{ij}$. Let the matrix B be given by

$$B = \begin{pmatrix} 0 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{pmatrix}.$$

It is clear that $\det(B) = 1$. Geometrically, we can think of the *parallelogram* constructed using the columns of A to lie inside the *parallelogram* constructed using the columns of B and as such, the area of *parallelogram* using columns of A is less than 1.

We next need to show that a_{ij} is strictly nonzero in the the range of integration. The integral under consideration is given by

$$a_{ij} = \int_0^x \int_0^x t^{i+b_1-1}(1-t)^{b_2} y^{j+b_1-1}(1-y)^{b_2} \operatorname{sgn}(y-t) dt dy.$$

The integral is being evaluated in the box, $[0, x] \times [0, x]$ for $0 < x < 1$. For $i \neq j$ (off-diagonal elements), the integrand is asymmetric around the diagonal joining $[0, 0]$ and $[x, x]$. However, for $i = j$, the integral is symmetric and due to the sign function, the integral is 0, which also follows from the skew symmetricity of the matrix. Hence, the integral is nonzero in the range of integration for every i and j. Therefore $0 < \det(A) < 1$. Hence, $0 < Pf(A) = \sqrt{\det(A)} < 1$. □

As in Subsection 3.2.2 the above approach can be extended to the cases of $\beta = 2, 4$ as well. In the general Jacobi ensemble setting, the extension of (3.24) that is needed is of the form

$$\Omega_\beta = \int_{[0,x]^p} \prod_{i=1}^{p} l_i^{c_1}(1-l_i)^{c_2} \prod_{i<j} |l_i - l_j|^\beta dl_1 dl_2 \ldots dl_p. \tag{3.35}$$

As in the general Gaussian ensemble setting, the only difference in the entire analysis is the treatment of the Vandermonde term. If $\beta = 2$ we can write $\prod_{i<j}(l_i - l_j)^2 = \det(x^i x^j)$ and apply (3.9), whereas for the case of $\beta = 4$, we apply (3.23) in (3.11).

3.4 Moments of the greatest root statistic for the bivariate case

Consider the $p = 2$ case. The joint distribution of the two eigenvalues of the multivariate beta distribution is given by

$$f(l_1, l_2) = \frac{1}{I(2, 1, a_1, a_2)} \prod_{i=1}^{2} l_i^{a_1 - \frac{1}{2} - 1} (1 - l_i)^{a_2 - \frac{1}{2} - 1} (l_1 - l_2)$$

where

$$I(2, 1, a_1, a_2) = \frac{\Gamma_2^{(1)}(2)}{\pi(\Gamma(\frac{3}{2}))^2} \frac{\Gamma_2^{(1)}(a_1)\Gamma_2^{(1)}(a_2)}{\Gamma_2^{(1)}(a_1 + a_2)}.$$

In the current section, we denote the normalising constant given above by C. Here, we evaluate the moments of the greatest root statistic when the number of dimensions is 2. For a 2×2 skew symmetric matrix, $A = (a_{ij})$, it is easy to see that $Pf(A) = a_{12}$, which normalised by the above constant gives the cdf of the greatest root. The support of the distribution of the greatest root is $(0, 1)$. We know that for a continuous positive random variable, X, the expected value can be written as

$$E[X] = \int_0^\infty P[X > x]dx.$$

We can thus write the expected value of the greatest root as

$$\begin{aligned} E[l_1] &= \int_0^1 P[l_1 > x]dx \\ &= \int_0^1 [1 - CPf(A)]dx \\ &= 1 - C \int_0^1 a_{12}dx \\ &= 1 - C \int_0^1 I_{12}^1 dx + C \int_0^1 I_{12}^2 dx. \end{aligned} \tag{3.36}$$

As in the previous section, we investigate the expectation term by term. Consider the first integral.

$$\begin{aligned}
E_1 &= \int_0^1 I_{12}^1 dx \\
&= \frac{2}{i+b_1}\sum_{n=0}^{b_2}\frac{(i+b_1)(-b_2)_n}{(i+b_1+n)(n!)}\int_0^1\int_0^x y^{\alpha+n-1}(1-y)^{1+b_2-1}dydx \\
&= \frac{2}{i+b_1}\sum_{n=0}^{b_2}\frac{(i+b_1)(-b_2)_n}{(i+b_1+n)(n!)}\int_0^1 B(\alpha+n-1,1+b_2,x)dx \\
&= \frac{2}{i+b_1}\sum_{n=0}^{b_2}\frac{(i+b_1)(-b_2)_n}{(i+b_1+n)(n!)}\int_0^1 \frac{x^{\alpha+n}}{(\alpha+n)}{}_2F_1(\alpha+n,-b_2,\alpha+n+1,x)dx. \qquad (3.37)
\end{aligned}$$

The following result gives the definite integral of a ${}_2F_1$ hypergeometric function in terms of a ${}_3F_2$ hypergeometric function

$$\int_0^1 x^{\rho-1}(1-x)^{\sigma-1}{}_2F_1(\alpha_1,\beta,\gamma,x)dx = \frac{\Gamma(\rho)\Gamma(\sigma)}{\Gamma(\rho+\sigma)}{}_3F_2(\alpha_1,\beta,\rho;\gamma,\rho+\sigma;1).$$

Setting $\rho=\alpha+n$ and $\sigma=1$ in (3.37), we get

$$\begin{aligned}
E_1 = \frac{2}{i+b_1}\sum_{n=0}^{b_2}&\frac{(i+b_1)(-b_2)_n}{(i+b_1+n)(n!)(\alpha+n)(\alpha+n+1)} \\
&{}_3F_2(\alpha+n,-b_2,\alpha+n+1;\alpha+n+1,\alpha+n+2;1). \qquad (3.38)
\end{aligned}$$

The second integral term in (3.36) is $E_2=\int_0^1 I_{12}^2 dx$ which is

$$\begin{aligned}
E_2 &= \Gamma^2(b_2+1)\sum_{n=0}^{b_2}\sum_{k=0}^{b_2}\frac{(-1)^n(-1)^k}{n!k!}\frac{1}{(1+b_1+n)\Gamma(b_2+1-n)} \\
&\qquad \frac{1}{(2+b_1+k)\Gamma(b_2+1-k)}\int_0^1 x^{3+2b_1+n+k}dx \\
&= \Gamma^2(b_2+1)\sum_{n=0}^{b_2}\sum_{k=0}^{b_2}\frac{(-1)^n(-1)^k}{n!k!}\frac{1}{(1+b_1+n)\Gamma(b_2+1-n)} \\
&\qquad \frac{1}{(2+b_1+k)\Gamma(b_2+1-k)}\frac{1}{(4+2b_1+n+k)}.
\end{aligned}$$

Thus, writing $E = 1 - CE_1 + CE_2$, we get an analytical expression for the expected value of the greatest root statistic when $p = 2$. Using exactly the same argument as above, analytical expressions of higher order moments can be easily derived. However, deriving moments for dimensions greater that 2 seems a much more difficult proposition.

3.5 Discussion

We have presented a unified analytical framework to derive the distribution function of the largest eigenvalue of the Gaussian orthogonal ensemble and the Jacobi ensemble exploiting the theory of special functions and the *Pfaffian* form. Extensions to the analytical evaluations for GUE/GSE as presented in Subsection (3.2.2) need to be carried out. Extensions to the theoretically interesting case of evaluating a similar analogue for GSE and the corresponding complex and quaternion cases in the double Wishart setting are also in the works. The next natural questions would be to provide analytical expressions for the moments (if any) of the respective random variables and also address the case of computing exact p-values and confidence intervals for the *greatest root statistic*. It would also be instructive to devise a method to simulate from the derived distributions.

A related but analytically different question to address is to understand the maximal domain of attraction of the Tracy-Widom distribution. In other words, if $X_1, X_2, \ldots, X_n$ are i.i.d. copies of Tracy-Widom distributed random variables, what is the distribution of $\max\{X_1, X_2, \ldots, X_n\}$? This question is the subject of Chapter 4.

Chapter 4

Domain of Attraction of Tracy-Widom Distribution

4.1 Introduction

The Tracy-Widom law appears as the limiting distribution of the largest eigenvalue of various random matrices. However, it cannot be parametrised as a classical extreme value distribution since the Tracy-Widom law arises as an asymptotic distribution for a dependent sequence of random variables, contrary to the limit distributions in extreme value theory. We investigate the maximum domain of attraction of an i.i.d. sequence of Tracy-Widom random variables and show that it belongs to the *Gumbel* domain of attraction. Classical analysis of the extremes of random variables have relied on the well developed extreme value theory. We don't present details related to extreme value theory here and refer the reader to [10] or [36] for details.

The finite sample exact distribution of the largest eigenvalue of many classes of random matries is hard to find. The computation of p-values corresponding to the largest eigenvalue arising in many multivariate analyses techniques involving a single or double Wishart setting are rather tedious, relying on cumbersome tables or specialised software. Good asymptotic approximations are therefore a dersirable goal. Johnstone in [24] showed that the limiting distribution of the largest eigenvalue of a Wishart random matrix has the Tracy-Widom distribution. More precisely, suppose $X = (X_{ij})_{n\times p}$ has entries that are i.i.d. $X_{ij} \sim N(0,1)$ and suppose the sample eigenvalues of the Wishart matrix X^TX are denoted by $l_1 > l_2 >$

$\ldots > l_p$. Then Johnstone [24] showed that,

$$\frac{l_1 - \mu_{np}}{\sigma_{np}} \rightarrow W_1 \sim F_1$$

where

$$\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2 \quad \text{and} \quad \sigma_{np} = (\sqrt{n-1} + \sqrt{p})(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}})^{\frac{1}{3}}.$$

and F_1 is the Tracy-Widom distribution corresponding to a Gaussian orthogonal ensemble, details of which are presented in Section 4.2.

The greatest root statistic, which is the largest eigenvalue of a multivariate beta distribution or a Jacobi ensemble, describes the null hypothesis distribution of several methods such as MANOVA, canonical correlations, testing equality of covariance matrices etc. The definition of a Jacobi ensemble and the greatest root statistic are presented in Chapter 3. Traditionally, calculation of the exact distribution of the greatest root statistic has once again relied on extensive tables or use of specialised software. Dumitriu et al. in [15], give an expression for the distribution function and the density function of the largest eigenvalue of a real Jacobi matrix as

$$P(l_1 < x) = C_{1,p} x^{\frac{pn_1}{2}} {}_2F_1(\frac{n_1}{2}, \frac{-n_2 + p + 1}{2}; \frac{n_1 + p + 1}{2}; xI), \tag{4.1}$$

where

$$C_{1,p} = \frac{\Gamma_p^{(1)}(\frac{n_1+n_2}{2})\Gamma_p^{(1)}(\frac{p+1}{2})}{\Gamma_p^{(1)}(\frac{n_1+p+1}{2})\Gamma_p^{(1)}(\frac{n_2}{2})}.$$

and ${}_2F_1(\cdot, \cdot; \cdot, xI)$ denotes the hypergeometric function with a matrix argument, which in this case is considered to be the identity matrix. Yet another spectacular universality result was proved by Johnstone in [25] where he showed that the logit transform of the greatest root statistic converges to the Tracy-Widom law corresponding to the Gaussian orthogonal

ensemble subject to a more complicated location and scale transformation. In hypothesis testing situation with the greatest root statistic, people have classically resorted to using a lower bound based on the F distribution, which is highly anti-conservative. In this regard, several examples motivated in Johnstone [26] provide sound evidence of using the Tracy-Widom approximation to compute corresponding p-values.

4.2 Tracy-Widom Distribution

A random matrix model is a probability space $(\Omega, \mathcal{F}, P)$ where the sample space Ω is a space of matrices and the probability measure P, is an invariant measure corresponding to the symmetricity of the matrix structure. The following two definitions are in order.

Definition 4.1. (GOE) *Consider matrices of the form,* $M_n = [X_{i,j}]_{i,j=1}^n$ *where* $X_{i,j} = X_{j,i} \sim N(0,1)$, $i < j$ *and* $X_{i,i} \sim \sqrt{2}N(0,1)$ *and are independent for all* (i,j). *Such a matrix is called a Gaussian Orthogonal Ensemble.*

Definition 4.2. (GUE) *Consider matrices of the form,* $M_n = [X_{i,j}]_{i,j=1}^n$ *where* $X_{i,j} = \overline{X}_{j,i}$, $X_{i,j} \sim N(0,1) + iN(0,1/2)$, $i < j$ *and* $X_{i,i} \sim N(0,1)$ *and are independent for all* (i,j). *Such a matrix is called a Gaussian Unitary Ensemble.*

The density function corresponding to the reference Lebesgue measure is given by

$$\frac{1}{Z_{n,\beta}} \exp(-\frac{\beta}{4}\text{tr}M^2)$$

where $\beta = 1$ corresponds to GOE and $\beta = 2$ corresponds to GUE. Analogously, one can define the density function for the Guassian symplectic ensmebles by setting $\beta = 4$. GSE is a Gaussian random matrix model defined over the field of quaternions.

For any matrix A in the above matrix ensembles, let $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_n := \lambda_{\max}$ denote the eigenvalues of A. Note that since we are dealing with symmetric and Hermitian matrices, all the eigenvalues are real. The joint density of the eigenvalues of both GOE and GUE, parametrised by β can be written as

$$f_\beta(x_1, x_2, \ldots, x_n) = C_{n,\beta} \prod_{1 \leq i \leq j \leq n} |x_i - x_j|^\beta \prod_{i=1}^{n} \exp(-\frac{\beta x_i^2}{2}) \tag{4.2}$$

The exact distribution of the largest eigenvalue from the joint distribution of all the eigenvalues in the right hand side of (4.2) involves a highly non-trivial integration. Moreover, eigenvalues of such random matrices were used to model very high energy levels of complex nuclei and this involved understanding the behaviour in an asymptotic setting. Prompted by the analytical difficulty of the integration and to address the physical question of interest, Tracy and Widom in their seminal works in [48, 49] derived the limit distribution, i.e., the Tracy-Widom distribution, of the largest eigenvalue arising in both GOE and GUE.

The Tracy-Widom cumulative distirbution function arising from the GUE is given by

$$F_2(x) = \exp\left(- \int_x^\infty (s - x) q^2(s) ds \right). \tag{4.3}$$

and the Tracy-Widom cumulative distribution function arising from the GOE is given by

$$F_1(x) = \exp\left(-\frac{1}{2} \int_0^\infty q(s) dy \right) (F_2((x))^{\frac{1}{2}}. \tag{4.4}$$

in terms of the solution $q(x)$ to classical Painlevé non-linear second order differential equation

$$q''(x) = xq(x) + 2q^3(x), \qquad q(x) \sim \mathrm{Ai}(x) \quad \text{as} \quad x \to \infty \tag{4.5}$$

where $\mathrm{Ai}(x)$ denotes the Airy function and where $q(x) \sim \mathrm{Ai}(x)$ means

$$\lim_{x \to \infty} \frac{q(x)}{\mathrm{Ai}(x)} = 1.$$

The above cumulative distribution functions can also be expressed as a Fredholm determinant but we don't present that representation here. More details can be found in any of [12, 24, 48, 49, 50]. The Tracy-Widom distribution is parameter-free and the approximate mean and variance corresponding to GOE ($\beta = 1$) and GUE ($\beta = 2$) are presented in Table 4.1.

Table 4.1 – TW Statistics

β	Mean	Variance
1	-1.21	1.607
2	-1.77	0.813

4.3 Domain of Attraction

We investigate the following problem. If $X_1, X_2, \ldots, X_n$ are i.i.d. copies of random variables having a Tracy-Widom distribution arising from the Gaussian Unitary Ensemble (GUE), then what domain of attraction does $M_n = \max\{X_1, X_2, \ldots, X_n\}$ belong to? We know from [50] that the eigenvalues of GUE are real. We state below the classical result, see for example [10], that characterises extremal distributions.

Theorem 4.1. (Fisher-Tippet Theorem)

Let (X_n) be a sequence of i.i.d. random variables. If there exist normalising constants $c_n > 0$ and $d_n \in \mathbb{R}$ and some non-degenerate distribution function H such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H$$

then H belongs to the type of one of the following three distribution functions:

1. *Frechet:* $\phi_\alpha(x) = \left\{ \begin{array}{ll} 0 & \text{for } x \leq 0 \\ \exp(-x^{-\alpha}) & \text{for } x > 0 \end{array} \right\}$

2. *Weibull:* $\psi_\alpha(x) = \left\{ \begin{array}{ll} \exp(-(-x)^\alpha) & \textit{for} \quad x \leq 0 \\ 1 & \textit{for} \quad x > 0 \end{array} \right\}$

3. *Gumbel:* $\Lambda(x) = \exp(-(\exp(-x))) \quad x \in \mathbb{R}.$

Theorem 4.2. (Fisher-Tippet (1928) and Gnedenko (1943)) *The class of extreme value distributions is $G_\gamma(ax+b)$ with $a > 0$, $b \in \mathbb{R}$ where*

$$G_\gamma(x) = \exp\left\{-(1+\gamma x)^{-1/\gamma}\right\}, \quad 1+\gamma x > 0$$

with $\gamma \in \mathbb{R}$ and where for $\gamma = 0$, the right hand side is interpreted as $\exp(-e^{-x})$. The parameter γ is called the extreme value index.

We have the following result from [10] that gives a sufficient condition for any continuous distribution to belong to a particular domain of attraction.

Theorem 4.3. *Let F be a distribution function and x^* be its right endpoint. Suppose $F''(x)$ exists and $F'(x)$ is positive for all x in some left neighbourhood of x^*. If*

$$\lim_{x \uparrow x^*} \frac{(1-F(x))F''(x)}{(F'(x))^2} = -\gamma - 1.$$

then F is in the domain of attraction of G_γ.

Theorem 4.4 states the result of the domain of attraction of the maximum of an i.i.d. sequence of random variables having the Tracy-Widom distribution arising from the GUE.

Theorem 4.4. *Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables having the Tracy-Widom distribution for the unitary case with cumulative distribution function F_2 as given in (4.3). Let $x^* = \sup\{x \in \mathbb{R} | F_2(x) < 1\}$ denote the right end point of F_2. Here $x^* = \infty$. Let $X = \max(X_1, X_2, \ldots, X_n)$ denote the maximum. Then, $F_2 \in D(G_0)$, i.e., F_2 belongs to the domain of attraction of the Gumbel Distribution.*

Before proving the theorem, we first prove a few lemmas that are required to prove the theorem.

Lemma 4.1. *Suppose $q(x)$ denotes the solution to the Painlevé II differential equation, and $q'(x)$ denotes the first derivative of $q(x)$, then for sufficiently large values of x, i.e., for $x \to \infty$,*

$$q'(x) \sim -x^{1/2} q(x).$$

Proof. It is well known that $q(x) \sim \mathrm{Ai}(x)$ as $x \to \infty$. Recall from [2] that the Airy function $\mathrm{Ai}(x)$ satisfies $\mathrm{Ai}''(x) = x\mathrm{Ai}(x)$ and as $x \to \infty$,

$$\mathrm{Ai}(x) \sim \frac{1}{2\sqrt{\pi} x^{1/4}} e^{-\frac{2}{3}x^{3/2}}.$$

Writing $\zeta = \frac{2}{3}x^{3/2}$, from Chapter 9 in [55], we see that

$$\mathrm{Ai}'(x) = -3^{1/6}\pi^{-1/2}\zeta^{4/3}e^{-\zeta}U(\frac{7}{6}, \frac{7}{3}, 2\zeta). \tag{4.6}$$

where $U(a, b, x)$ is a standard solution to the Kummer's equation

$$x\frac{d^2w}{dx^2} + (b - x)\frac{dw}{dx} - aw = 0.$$

and is uniquely determined by the property that

$$U(a, b, x) \sim x^{-a} \quad \text{as} \quad x \to \infty$$

Thus as $x \to \infty$, the function $U(\cdot, \cdot, \cdot)$ in (4.6) becomes

$$U(\frac{7}{6}, \frac{7}{3}, 2\zeta) \sim (2\zeta)^{-7/6}. \tag{4.7}$$

Hence, asymptotically, (4.6) is

$$\mathrm{Ai}'(x) \sim -3^{1/6}\pi^{-1/2}\zeta^{4/3}e^{-\zeta}(2\zeta)^{-7/6}.$$

which simplifies to

$$\begin{aligned}\mathrm{Ai}^{'}(x) &\sim -\frac{1}{2\sqrt{\pi}}x^{1/4}e^{-\frac{2}{3}x^{3/2}} \\ &\sim -x^{1/2}\mathrm{Ai}(x). \end{aligned} \tag{4.8}$$

From [41], we see that $q(x)$ satisfies the boundary condition that $q(x) \to 0$ as $x \to \infty$ and from [55], we see that the Airy function is such that $\lim_{x\to\infty} \mathrm{Ai}(x) = 0$. Consequently, by l'Hospital rule, $q'(x) \sim \mathrm{Ai}'(x)$ as $x \to \infty$. Thus, from (4.8) we see

$$q^{'}(x) \sim -x^{1/2}q(x) \quad \text{as} \quad x \to \infty.$$

□

Lemma 4.2. *Let $F_2^{'}(x)$ and $F_2^{''}(x)$ denote the first and second derivative of $F_2(x)$, the cumulative distribution of the Tracy-Widom distribution arising from GUE ($\beta = 2$). Then,*

$$\begin{aligned} F_2^{'}(x) &= F_2(x)R(x). \\ \frac{F_2^{''}(x)}{F_2^{'}(x)} &= R(x) + \frac{R'(x)}{R(x)}. \end{aligned}$$

where

$$R(x) = \int_x^\infty q^2(s)ds. \tag{4.9}$$

Proof. Let

$$\phi(x) = \int_x^\infty (s-x)q^2(s)ds = \int_x^\infty sq^2(s)ds - xR(x). \tag{4.10}$$

When $s \to \infty$, then $sq^2(s) \to 0$ because $q(s)$ behaves asymptotically like the Airy function for sufficiently large values of s. Thus, letting the upper limit of the integration on the right hand side of (4.10) be some x_0 where x_0 is "sufficiently" large, we can apply the fundamental

theorem of calculus to get

$$\frac{d}{dx}\left[\int_x^{x_0} sq^2(s)ds\right] = -xq^2(x) \quad \text{for} \quad x_0 \to \infty. \tag{4.11}$$

Similar argument gives

$$R'(x) = -q^2(x). \tag{4.12}$$

Combining (4.11) and (4.12) we get the derivative of $\phi(x)$ as

$$\phi'(x) = -R(x). \tag{4.13}$$

Rewriting (4.3) as $F_2(x) = \exp(-\phi(x))$ and then taking logarithms on both sides and differentiating with respect to x we get,

$$\frac{F_2'(x)}{F_2(x)} = -\phi'(x) = R(x) \tag{4.14}$$

which implies

$$F_2'(x) = F_2(x)R(x).$$

Taking logarithms and differentiating again with respect to x in the above equation we get,

$$\frac{F_2''(x)}{F_2'(x)} = \frac{F_2'(x)}{F_2(x)} + \frac{R'(x)}{R(x)} = R(x) + \frac{R'(x)}{R(x)}. \tag{4.15}$$

□

Lemma 4.3. *For $x \to \infty$, $R(x)$ is asymptotically given by*

$$R(x) = \exp\Big(-\frac{4}{3}x^{3/2}\Big)\Big(\frac{1}{8\pi x} + O(x^{-5/2})\Big). \tag{4.16}$$

Proof. $R(x)$ is as in (4.9). From Baik et al. [3], we get an equivalent representation of $R(x)$ as

$$R(x) = [q'(x)]^2 - xq^2(x) - q^4(x). \tag{4.17}$$

Bassom et al. in [1] provide asymptotic expansions for $q(x)$ and $q'(x)$ as

$$q(x) = \frac{1}{2\sqrt{\pi}} x^{-1/4} \exp\big(-\frac{2}{3}x^{3/2}\big)\big[1 - \frac{5}{48}x^{-3/2} + O(x^{-3})\big]. \tag{4.18}$$

$$q'(x) = -\frac{1}{2\sqrt{\pi}} x^{1/4} \exp\big(-\frac{2}{3}x^{3/2}\big)\big[1 + \frac{7}{48}x^{-3/2} + O(x^{-3})\big]. \tag{4.19}$$

From the above equations we get,

$$\begin{aligned} xq^2(x) &= \frac{1}{4\pi} x^{1/2} \exp\big(-\frac{4}{3}x^{3/2}\big)\big[1 - \frac{5}{48}x^{-3/2} + O(x^{-3})\big]^2 \\ &= \frac{1}{4\pi} x^{1/2} \exp\big(-\frac{4}{3}x^{3/2}\big)\big[1 - \frac{5}{24}x^{-3/2} + O(x^{-3})\big] \\ &= \frac{1}{4\pi} \exp\big(-\frac{4}{3}x^{3/2}\big)\big[x^{1/2} - \frac{5}{24x} + O(x^{-5/2})\big]. \end{aligned} \tag{4.20}$$

and

$$\begin{aligned} [q'(x)]^2 &= \frac{1}{4\pi} x^{1/2} \exp\big(-\frac{4}{3}x^{3/2}\big)\big[1 + \frac{7}{48}x^{-3/2} + O(x^{-3})\big]^2 \\ &= \frac{1}{4\pi} x^{1/2} \exp\big(-\frac{4}{3}x^{3/2}\big)\big[1 + \frac{7}{24}x^{-3/2} + O(x^{-3})\big] \\ &= \frac{1}{4\pi} \exp\big(-\frac{4}{3}x^{3/2}\big)\big[x^{1/2} + \frac{7}{24x} + O(x^{-5/2})\big]. \end{aligned} \tag{4.21}$$

Expanding $q^4(x)$ similarly and combining the expressions obtained in (4.20) and (4.21) we get the required result. □

PROOF OF THEOREM

Proof. We use Theorem 4.3 to prove the result. In other words, we need to to show that

$$\lim_{x\to\infty} \frac{[1 - F_2(x)]F_2''(x)}{[F_2'(x)]^2} = -1. \tag{4.22}$$

If the limit in (4.22) evaluates to -1, then the extreme value index $\gamma = 0$ which gives a sufficient condition that the Tracy-Widom distribution for GUE belongs to the Gumbel domain of attraction. The left hand side of (4.22) can be written as

$$L = \left[\frac{1 - F_2(x)}{F_2'(x)}\right] \cdot \left[\frac{F_2''(x)}{F_2'(x)}\right].$$

From Lemma 4.2 we get

$$L = \left[\frac{1-F_2(x)}{F_2(x)R(x)}\right]\cdot\left[R(x)+\frac{R'(x)}{R(x)}\right]$$
$$= [1-F_2(x)] + \left[\frac{1-F_2(x)}{F_2(x)}\right]\cdot\left[\frac{R'(x)}{R^2(x)}\right]. \tag{4.23}$$

Since $F_2(x)$ is a cdf, $\lim_{x\to\infty}[1-F_2(x)] = 0$. We thus need to show that the second term in the right hand side of (4.23) goes to -1 as $x \to \infty$. Since $F_2(x) = \exp[-\phi(x)]$ and using (4.12), we can express the second term in (4.23) as

$$L_2 = -\lim_{x\to\infty}\frac{q^2(x)[e^{\phi(x)}-1]}{R^2(x)}. \tag{4.24}$$

Let $J(x) = \frac{1}{2\sqrt{\pi}}x^{-1/4}e^{-\frac{2}{3}x^{3/2}}$. It is well known that asymptotically $\mathrm{Ai}(x) \sim J(x)$ as $x \to \infty$, see for example [55]. Thus, $q(x) \sim J(x)$. Hence,

$$L_2 = -\lim_{x\to\infty}\frac{J^2(x)[e^{\phi(x)}-1]}{R^2(x)}.$$

Applying l'Hosptial rule we get,

$$L_2 = -\lim_{x\to\infty}\frac{J^2(x)e^{\phi(x)}\phi'(x) + 2J(x)J'(x)[e^{\phi(x)}-1]}{2R(x)R'(x)}.$$

It is easily seen that

$$J'(x) = -\left[\frac{1}{4x} + x^{1/2}\right]J(x).$$

It follows that $J'(x) \sim -\sqrt{x}J(x)$. Thus,

$$2J(x)J'(x)[e^{\phi(x)}-1] = 2J(x)[e^{\phi(x)}-1]\left[\frac{J'(x)}{-\sqrt{x}J(x)}\right][-\sqrt{x}J(x)]$$
$$= -2\sqrt{x}J^2(x)[e^{\phi(x)}-1]\left[\frac{J'(x)}{-\sqrt{x}J(x)}\right].$$

Thus, (4.24) is

$$\begin{aligned} L_2 &= -\lim_{x\to\infty} \frac{J^2(x)R(x)e^{\phi(x)} + 2\sqrt{x}J^2(x)[e^{\phi(x)}-1]\big[J'(x)/-\sqrt{x}J(x)\big]}{2q^2(x)R(x)} \\ &= -\left[\lim_{x\to\infty}\frac{1}{2}\cdot\frac{J^2(x)e^{\phi(x)}}{q^2(x)} + \lim_{x\to\infty}\frac{\sqrt{x}J^2(x)[e^{\phi(x)}-1]\big[J'(x)/-\sqrt{x}J(x)\big]}{q^2(x)R(x)}\right] \\ &= -\left[\lim_{x\to\infty}\frac{1}{2}\cdot\frac{e^{\phi(x)}}{q^2(x)/J^2(x)} + \lim_{x\to\infty}\frac{\sqrt{x}[e^{\phi(x)}-1]}{R(x)}\cdot\frac{\big[J'(x)/-\sqrt{x}J(x)\big]}{q^2(x)/J^2(x)}\right]. \end{aligned} \tag{4.25}$$

Observe that $\lim_{x\to\infty}\phi(x)=0$, so $\lim_{x\to\infty}e^{\phi(x)}=1$ and $\lim_{x\to\infty}q^2(x)/J^2(x)=\lim_{x\to\infty}\big[J'(x)/-\sqrt{x}J(x)\big]=1$. Thus, if it is shown that $\lim_{x\to\infty}\sqrt{x}[e^{\phi(x)}-1]/R(x)=1/2$, we get the required result.

$$\begin{aligned} \lim_{x\to\infty}\frac{\sqrt{x}[e^{\phi(x)}-1]}{R(x)} &= \lim_{x\to\infty}\frac{\sqrt{x}(1-e^{-\phi(x)})}{e^{-\phi(x)}R(x)} \\ &= \lim_{x\to\infty}\frac{\sqrt{x}(1-F_2(x))}{F_2(x)}\cdot\frac{1}{R(x)}. \end{aligned} \tag{4.26}$$

Observe that $F_2(x)=F(x)^2$ where

$$F(x)=\exp\left(-\frac{1}{2}\int_x^\infty (s-x)q^2(s)ds\right).$$

From [50] we note that

$$F(x)=1-\frac{e^{-\frac{4}{3}x^{3/2}}}{32\pi x^{3/2}}\Big(1+O(x^{-3/2}\Big).$$

Thus, asymptotically,

$$F_2(x)^{-1}=F(x)^{-2}=1+\frac{e^{-\frac{4}{3}x^{3/2}}}{16\pi x^{3/2}}\Big(1+O(x^{-3/2})\Big).$$

and,

$$(1-F_2(x))=\frac{e^{-\frac{4}{3}x^{3/2}}}{16\pi x^{3/2}}\Big(1+O(x^{-3/2})\Big).$$

Hence we get the following asymptotic expression as

$$\frac{1-F_2(x)}{F_2(x)}=\left(\frac{1}{16\pi x^{3/2}}+O(x^{-3})\right)e^{-\frac{4}{3}x^{3/2}}. \tag{4.27}$$

Using Lemma 4.3 and (4.27), the right hand side of (4.26) has the following asymptotic expression

$$\begin{aligned}\frac{\sqrt{x}(1-F_2(x))}{F_2(x)}\cdot\frac{1}{R(x)} &= \frac{\sqrt{x}\Big(\frac{1}{16\pi x^{3/2}}+O(x^{-3})\Big)e^{-\frac{4}{3}x^{3/2}}}{\Big(\frac{1}{8\pi x}+O(x^{-5/2})\Big)e^{-\frac{4}{3}x^{3/2}}}\\ &= \frac{\frac{1}{16\pi x}+O(x^{-7/2})}{\frac{1}{8\pi x}+O(x^{-5/2})}\\ &= \frac{1}{2}\cdot\frac{1}{1+O(x^{-3/2})}+\frac{O(x^{-5/2})}{1+O(x^{-3/2})} \end{aligned} \tag{4.28}$$

The right hand side in (4.28) goes to $1/2$ as $x\to\infty$. As such, (4.25) is

$$\begin{aligned}L_2 &= -\Bigg[\lim_{x\to\infty}\frac{1}{2}\cdot\frac{e^{\phi(x)}}{q^2(x)/J^2(x)}+\lim_{x\to\infty}\frac{\sqrt{x}[e^{\phi(x)}-1]}{R(x)}\cdot\lim_{x\to\infty}\frac{\big[J'(x)/-\sqrt{x}J(x)\big]}{q^2(x)/J^2(x)}\\ &= -\big[\frac{1}{2}\cdot 1+\frac{1}{2}\cdot 1\big]\\ &= -1\end{aligned}$$

which is the required result. □

In Figure 4.1, we plot the expression in the left hand side of (4.22) for the Tracy-Widom distribution corresponding to GUE ($\beta=2$) and GOE ($\beta=1$) for a grid of x values. Analytical derivation of the limit result for the GOE case seems more complicated than the GUE case. However, the convergence of the limit to -1 can be gleaned from the figure for both the cases which is encouraging since the tail behaviour of the distribution for the GOE case seems to indicate that it should belong to the Gumbel domain of attraction.

4.4 Simulation

We conducted a basic simulation study to compare the behaviour of the maximum of i.i.d. Tracy-Widom random variables corresponding to $\beta=2$ and the Gumbel distribution. In

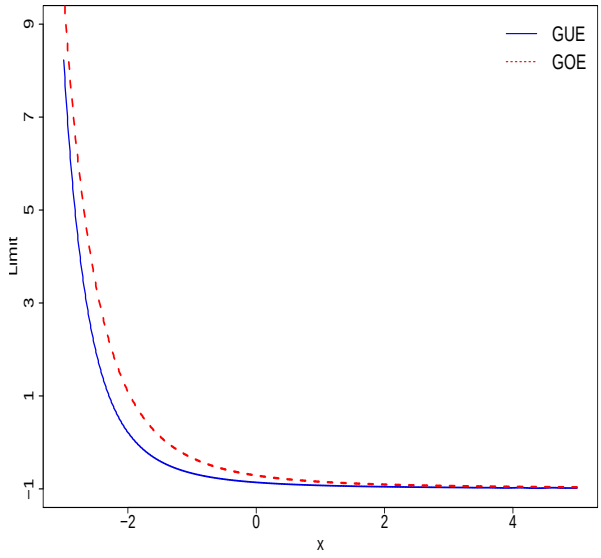

Figure 4.1 – Limit corresponding to GUE and GOE

each simulation run, generate an i.i.d. sequence, $X_1, X_2, \ldots, X_k$, of Tracy-Widom random variables of length $k = 10000$ and store the value of $M = \max\{X_1, X_2, \ldots, X_k\}$ from each run. This exercise is repeated $n = 1000$ times to get a distribution of M, the maximum of i.i.d. Tracy-Widom random variables. In Table 4.2, we present some basic empirical statistics of the simulated distribution of maximums and also report the corresponding theoretical statistics from the Gumbel distribution. It can be seen that the basic measures of mean and standard deviation of the simulated distribution of maximums and the Gumbel distribution are indeed very close to each other. In Table 4.3, we report the $1^{\text{st}}, 25^{\text{th}}, 75^{\text{th}}, 90^{\text{th}}, 95^{\text{th}}$, and 99^{th} quantiles of the simulated maximums and the theoretical Gumbel distribution respectively. It can be clearly seen that even at the $90^{\text{th}}, 95^{\text{th}}$ and 99^{th} quantile, the corresponding numbers are very close. In fact, for any hypothesis test conducted at $\alpha = 0.05$ or even $\alpha = 0.01$, using the Tracy-Widom distribution, the p-value computations using a Gumbel approximation would be conservative.

In Figure 4.2, we present the QQ-plot of the simulated maximums and the Gumbel distribution. We observe a departure on the extreme right tail. However, this happens well beyond the 99^{th} quantile and for all practical purposes, this would not cause any problem in carrying out any hypothesis tests. In Figure 4.3, we present a histogram of the simulated maximum of Tracy-Widom random variables overlaid with the theoretical Gumbel density. It can be seen that the fit seems to be rather good. The exercise was done on the R platform with the newly available RMTSTAT software. The location term for normalisation is $b_n = U(n)$ where $U(n)$ is the left continuous inverse of $1/(1 - F_2)$. $U(n)$ can be treated as the $100 * (1 - 1/n)$ quantile of the distribution, which can be easily obtained using the the quantile function of the software. The scaling term is $a_n = 1/(nF_2'(b_n))$. Since the cumulative distribution function is continuous, $F_2'(b_n)$ is just the density function evaluated at b_n, which can also be easily obtained using the package RMTSTAT.

Table 4.2 – Comparison of some statistics of simulated max TW with true Gumbel

Measures	Mean	S.D.
Max TW	0.5771	1.2586
Gumbel	0.5572	1.2825

Table 4.3 – Comparison of some Quantiles of simulated max TW and Gumbel

Quantiles	1^{st}	25^{th}	75^{th}	90^{th}	95^{th}	99^{th}
Max TW	-2.1479	-0.3229	1.2445	2.2815	2.9614	4.2539
Gumbel	-2.1439	-0.3140	1.2114	2.2160	2.9212	4.5204

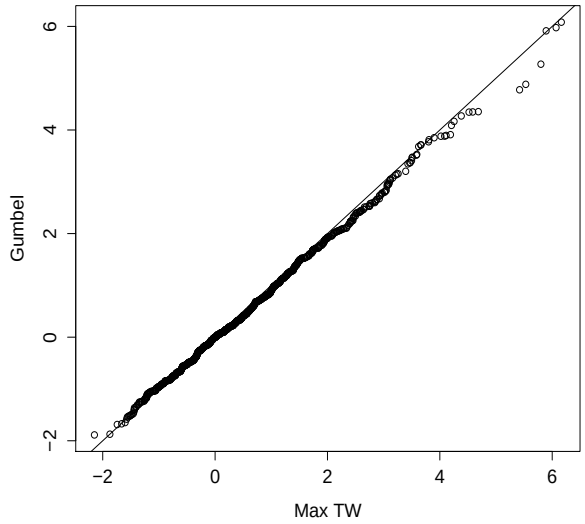

Figure 4.2 – QQ plot of Max TW (GUE) with theoretical Gumbel

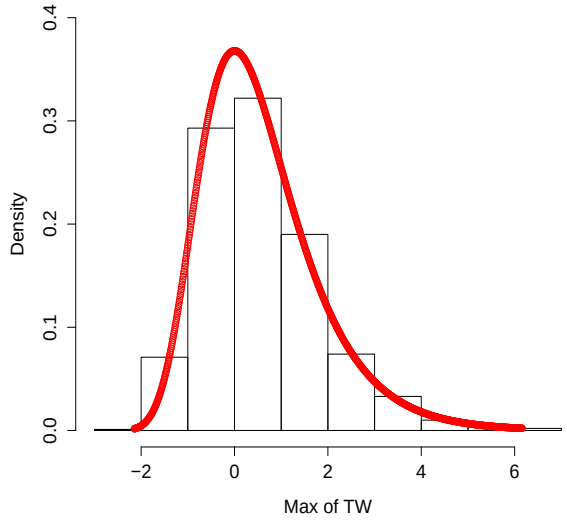

Figure 4.3 – Histogram of Max TW GUE overlaid with the Gumbel density

4.5 Statistical Applications

Suppose there are n_1 i.i.d. samples from $\mathbf{x} \sim N(\mu_1, \Sigma_1)$ and n_2 i.i.d. samples from $\mathbf{y} \sim N(\mu_2, \Sigma_2)$. Further suppose the objective is to compare the equality of the two covariance matrices, i.e., define the null and the alternative hypothesis as $H_0 : \Sigma_1 = \Sigma_2$ vs $H_a : \Sigma_1 \neq \Sigma_2$. Let S_i be the covariance estimates which are independent Wishart distributed matrices with n_i degrees of freedom, i.e., $n_i S_i \sim W_p(n_i, \Sigma_i)$ for $i = 1, 2$. Then, a test of the null hypothesis is based on the largest eigenvalue, λ_1 of $(n_1 S_1 + n_2 S_2)^{-1} n_2 S_2$, which under the null hypothesis has the $\lambda_1(p, n_1, n_2)$ distribution. It can be seen in Johnstone [26] that the Tracy-Widom distribution corresponding to $\beta = 1$ provides very good approximation to the distribution of $\lambda_1(p, n_1, n_2)$ to test the null hypothesis.

Consider the following *union-intersection* type sequence of hypotheses to be conducted. Let

$$H_{01} : \Sigma_{11} = \Sigma_{12},\ H_{02} : \Sigma_{21} = \Sigma_{22},\ \ldots,\ H_{0k} : \Sigma_{k1} = \Sigma_{k2}.$$

Define the multivariate hypothesis H_0 as $H_0 = \cap H_{0j}$ for $j = 1, 2, \ldots, k$. This implies that H_0 is true if and only if each of the component hypothesis H_{0j} is true. Thus, *accept* H_0 if and only if every component hypothesis H_{0j} is *accepted*. We can equivalently say that we reject H_0 if any component hypothesis H_{0j} is rejected. Let R_j denote the rejection region corresponding to the j^{th} hypothesis test. Then by the union-intersection test principle, $R = \cup R_j$ denotes the rejection region corresponding to H_0. Let n_{11}, n_{12} denote the sample sizes corresponding to the hypothesis test H_{01}. Let n_{21}, n_{22} be the sample sizes corresponding to the hypothesis test H_{02}. In general, let n_{j1}, n_{j2} denote the sample sizes for the j^{th} hypothesis test for $j = 1, 2, \ldots, k$. Let S_{j1}, S_{j2} denote the covariance estimators for the j^{th} hypothesis test. Thus, the test statistic for H_{0j} is $\lambda_{1j}(p, n_{j1}, n_{j2})$, which is the largest eigenvalue of

$(n_{j1}S_{j1} + n_{j2}S_{j2})^{-1}n_{j2}S_{j2}$. For each of the component hypothesis, let T_j denote the Tracy-Widom approximation of $\lambda_{1j}(p, n_{j1}, n_{j2})$. Thus, $R_j = \{T_j > c_j\}$ where c_j is a constant. Let $M_k = \max\{T_1, T_2, \ldots, T_k\}$ and $R = \cup R_j$. Therefore, reject H_0 if M_k falls in the rejection region, i.e., $R = \{M_k > c\}$ for some constant c which is chosen so that $P_{H_0}(M_k > c) = \alpha$. For large k, $M_k \sim G$ where G has a Gumbel distribution. Hence, in such *union-intersection* type test constructions for a large number of independent tests where the greatest root statistic is used, one could use a Gumbel distribution approximation to compute approximate p-values.

BIBLIOGRAPHY

[1] A.Bassom, P.A.Clarkson, C.K.Law, and McLeod, J. Application of Uniform Asymptotics to the Second Painlevé Transcendent. *Archive for Rotational Mechanics and Analysis 143* (1998), 241–271.

[2] Abramowitz, M., and Stegun, I. A. *Handbook of Mathematical Functions*. Dover Publications, New York, 1965.

[3] Baik, J., Buckingham, R., and DiFranco, J. Asymptotics of Tracy-Widom Distributions and the Total Integral of a Painlevé II Function. *Communications in Mathematical Physics 280(2)* (2008), 463–497.

[4] Bartlett, P., Boucheron, S., and Lugosi, G. Model Selection and Error Estimation. *Machine Learning 48* (2002), 85–113.

[5] Bock, M. Employing Vague Inequality Information in the Estimation of Normal Mean Vectors. *Technical Report* (1982).

[6] Bruijn, N. D. On Some Multiple Integrals Involving Determinants. *Journal of the Indian Mathematical Society 19* (1955), 133–151.

[7] Candes, E., and Tao, T. Decoding by Linear Programming. *IEEE Transactions on Information Theory 51*, 12 (Dec 2005), 4203–4215.

[8] Chatterjee, A., and Lahiri, S. Bootstrapping Lasso Estimators. *Manuscript* (2010).

[9] Constantine, A. Some Non-Central Distribution Problems arising in Multivariate Analysis. *Annals of Mathematical Statistics 34* (1963), 1270–1285.

[10] de Haan, L., and Ferreira, A. *Extreme Value Theory: An Introduction*. Springer Verlag, 2006.

[11] Deutsch, F. *Best Approximation in Inner Product Spaces*. Springer Verlag, 2001.

[12] Dieng, M., and Tracy, C. Applications of Random Matrix Theory to Multivariate Statistics. *arXiv.math/0603543v1* (2006).

[13] Donoho, D. Minimax Risk over l_p balls for l_q Error. *Probability Theory and Related Fields 99* (1994), 277–303.

[14] Donoho, D., and Johnstone, I. Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of American Statistical Association 90* (1995), 1200–1224.

[15] Dumitriu, I., and Koev, P. Distributions of the Extreme Eigenvlaues of Beta-Jacobi Random Matrices. *SIAM Journal of Matrix Analysis 30* (2008), 1–6.

[16] Efron, B. The Estimation of Prediction Error: Covariance Penalties and Cross Validation. *Journal of the American Statistical Association 99* (2004).

[17] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least Angle Regression. *Annals of Statistics 32*, 2 (2004), 407–499.

[18] Fourdrinier, D., and Lepelletier, P. Estimating a General Function of a Quadratic Function. *Annals of the Institute of Statistical Mathematics 60* (2006), 85–119.

[19] Fourdrinier, D., Strawderman, W. E., and Wells, M. T. Robust Shrinkage Estimation for Elliptically Symmetric Distributions with Unknown Covariance Matrix. *Journal of Multivariate Analysis 85*, 1 (April 2003), 24–39.

[20] Fourdrinier, D., and Wells, M. T. Comparaisons de procedures de selection dun model de regression: Une approache decisionnelle. *C.R. Acad. Sci. Paris Serie I 319* (1994), 865–870.

[21] Fourdrinier, D., and Wells, M. T. Estimation of a Loss Function for Spherically Symmetric Distributions in the General Linear Model. *Annals of Statistics 23(2)* (1995), 571–592.

[22] James, W. D., and Stein, C. Estimation With Quadratic Loss. *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability 1* (1961), 361–379.

[23] Johnstone, I. Inadmissibility of Unbiased Estimates of Loss. In *Statistical Decision Theory and Related Topics*, vol. 1. Springer-Verlag, New York, 1988, pp. 361–379. Shanti. S. Gupta and James. O. Berger (Editors).

[24] Johnstone, I. On the Distribution of the Largest Eigenvalue in Principal Component Analysis. *The Annals of Statistics 29(2)* (2001), 295–327.

[25] Johnstone, I. Multivariate Analysis and Jacobi Ensembles:Largest Eigenvalue, Tracy-Widom Limits and Rates of Convergence. *Annals of Statistics 36(6)* (2008), 2638–2716.

[26] Johnstone, I. Approximate Null Distribution of the Largest Root in Multivariate Aanalysis. *The Annals of Applied Statistics 3(4)* (2009), 1616–1633.

[27] Kato, K. On the Degrees of Freedom in Shrinkage Estimation. *Journal of Multivariate Analysis 100*, 7 (2009), 1338–1352.

[28] Knight, K., and Fu, W. Asymptotics for Lasso Type Estimators. *Annals of Statistics 285* (2000), 1356–1378.

[29] Koev, P., and Edelman, A. The Efficient Evaluation of the Hypergeometric Function of a Matrix Argument. *Mathematics of Computation 75(254)* (2006), 833–846.

[30] Kuriki, S., and Takemura, A. Shrinkage Estimation towards a Closed Convex Set with a Smooth Boundary. *Journal of Multivariate Analysis 75*, 1 (2000), 79–111.

[31] Lele, C. Admissibility Results in Loss Estimation. *Annals of Statistics 21(1)* (1993), 378–390.

[32] Marčenko, V., and Pastur, L. Distributions of Eigenvalues of Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik 1(4)* (1967), 507–536.

[33] Mehta, M. L. *Random Matrices.* Academic Press, 1990.

[34] Muirhead, R. J. *Aspects of Multivariate Statistical Theory.* Wiley Series in Probability and Statistics, 1982.

[35] Osborne, M. R., Presnell, B., and Turlach, B. A. On the Lasso and its Dual. *Journal of Computational and Graphical Statistics 92* (2000), 319–337.

[36] Resnick, S. I. *Extreme Values, Regular Variation and Point Processes.* Springer Series in Operations Research and Financial Engineering, New York, 2008.

[37] Rukhin, A. L. Estimated Loss and Admissible Loss. In *Statistical Decision Theory and Related Topics*, vol. 1. Springer-Verlag, New York, 1988, pp. 409–418. Shanti. S. Gupta and James. O. Berger (Editors).

[38] Sandved, E. Ancillary Statistics and Estimation of the Loss in Estimation Problems. *Annals of Mathematical Statistics 39(5)* (1968), 1755–1758.

[39] Selberg, A. Remarks on a Multiple Integral. *Norsk Mat. Tidsskr. 26* (1944), 71–78.

[40] Sengupta, D., and Sen, P. K. Shrinkage Estimation in a Restricted Parameter Space. *Sankhya: The Indian Journal of Statistics Series A 53*, 3 (Oct 1991), 389–411.

[41] S.P.Hastings and J.B. McLeod, TITLE = A Boundary Value Problem Associated with the Second Painlevé Transcendent and the Korteweg-de Vries Equation, J. . A. y. . . v. . . p. . .

[42] Srivastava, M. Singular Wishart and Multivariate Beta Distributions. *Annals of Statistics 31(5)* (2003), 1537–1560.

[43] Stein, C. M. Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics 9*, 1 (1981), 1135–1151.

[44] Stoer, J., and Witzgall, C. *Convexity and Optimization in Finite Dimensions I.* Springer-Verlag, 1970.

[45] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B 58*, 1 (1996), 267–288.

[46] Tracy, C., and Widom, H. Level Spacing Distributions and the Airy Kernel. *Physics Letters B(1-2)* (1993), 115–118.

[47] Tracy, C., and Widom, H. Level Spacing Distributions and the Airy Kernel. *Communications in Mathematical Physics 159(1)* (1994), 151–174.

[48] Tracy, C., and Widom, H. On Orthogonal and Symplectic Matrix Ensembles. *Communications in Mathematical Physics 177(3)* (1996), 727–754.

[49] Tracy, C., and Widom, H. Correlation Functions, Cluster Functions and Spacing Distributions for Random Matrices. *Journal of Statistical Physics 92* (1998), 809–835.

[50] Tracy, C., and Widom, H. The Distributions of Random matrix Theory and their Applications. *Stanford Institute for Theoretical Economics* (2008).

[51] Wainwright, M. Sharp thresholds for High Dimensional and Noisy Recovery of Sparsity. *arXiv:math.ST/0605740* (2006).

[52] Wan, A. T., and Zou, G. On Unbiased and Improved Loss Estimation for the Mean of a Multivariate Normal Distribution with Unknown Variance. *Journal of Statistical Planning and Inference 119* (2004), 17–22.

[53] Webster, R. *Convexity*. Oxford University Press, 1994.

[54] Wigner, E. P. On the Distribution of the Roots of Certain Symmetric Matrices. *Annals of Mathematics 67(2)* (1958), 325–327.

[55] W.J.Olver, F. *Digital Library of Mathematical Functions*. NIST Digital Library of Mathematical Functions, 2012.

[56] W.Magnus, F.Oberhettinger, and Soni, R. *Formulas and Theorems for the Special Functions of Mathematical Physics*. Springer, 1966.

[57] Zhao, P., and Yu, B. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research 7* (2006), 2541–2563.

[58] Zou, H., Hastie, T., and Tibshirani, R. On the Degrees of Freedom of the Lasso. *The Annals of Statistics 35*, 5 (2007), 2173–2192.